

Convex Relaxation for Community Detection with Covariates

Bowei Yan, Purnamrita Sarkar
University of Texas at Austin

Abstract

Community detection in networks is an important problem in many applied areas. In this paper, we investigate this in the presence of node covariates. Recently, an emerging body of theoretical work has been focused on leveraging information from both the edges in the network and the node covariates to infer community memberships. However, so far the role of the network and that of the covariates have not been examined closely. In essence, in most parameter regimes, one of the sources of information provides enough information to infer the hidden cluster labels, thereby making the other source redundant. To our knowledge, this is the first work which shows that when the network and the covariates carry “orthogonal” pieces of information about the cluster memberships, one can get improved clustering accuracy by using them both, even if each of them fails individually.

1 Introduction

Community detection in networks is a fundamental problem in machine learning and statistics. A variety of important practical problems like analyzing socio-political ties among leading politicians [15], understanding brain graphs arising from diffusion MRI data [5], investigating ecological relationships between different tiers of the food chain [21] can be framed as community detection problems. Much attention has been focused on developing models and methodology to recover latent community memberships. Among generative models, the stochastic block model [19] and its variants [16] have attracted a lot of attention, since their simplicity facilitates efficient algorithms and asymptotic analysis [10, 30].

Although most real world network datasets come with covariate information associated with nodes and pairs of nodes, existing approaches are primarily focused on using the network for inferring the hidden community memberships or labels. Recently there have been some works towards joint inference which aim at combining the information from the network with that from the covariates [5, 18, 38]. Theoretically the main challenge is to separate the effect of the covariates from that of the network. As we will describe in detail, in a broad regime of parameters, many methods which only use networks can yield asymptotically consistent estimates of the unknown labels. In this respect, the balance is rather delicate because, either the network has enough signal to infer the hidden labels, or the covariates do. To our knowledge, existing theoretical works do not make a clear distinction between the information from the two

sources. In contrast, we show that when the signals from the two sources are in some sense orthogonal to each other, under a broad parameter regime, the two sources can be combined to yield good clustering accuracy despite the individual failure of either method.

Take for example the Mexican political elites network (described in detail in Section 4). This dataset comprises of 35 politicians (military or civilian) and their connections. The associated covariate for each politician is the year when he/she came into power. After the revolution the political arena was dominated by the military who have gradually been replaced by the civilians. Hence those who came into power later are more likely to be civilians. We found two politicians who had an equal number of connections to the military group and civilian groups and hence are hard to be classified into either group just from the network. Here the temporal covariate is crucial in resolving which group they belong to. Our method can successfully classify these politicians unlike other existing methods.

Typically in asymptotic analysis of networks, one lets the linkage probabilities go to zero in order to allow for sparse graphs. This rate essentially puts different growth conditions on the average degree of the network. We will denote the regime where the average degree grows faster than logarithm of the number of nodes n as the dense case and the regime where the average degree is constant as the sparse case. Most network-based community detection schemes like Spectral Clustering [30], likelihood and modularity based methods [1, 4] and semidefinite relaxations [2, 9, 10] give perfect classification in the limit in the dense case.

The sparse case is harder to analyze and has recently attracted a lot of attention. It has been shown that regularized Spectral Clustering [23] and semidefinite relaxations (SDP) of likelihood based methods [17] can achieve near perfect clustering, i.e. a small albeit constant fraction of nodes are misclassified. Most of these works assume strong assortativity, i.e. the smallest within cluster linkage probability is larger than the largest across cluster linkage probability. Recently Amini and Levina [2] have proposed SDP relaxations to achieve perfect clustering in the dense case for weakly assortative settings. A model is defined to be weakly assortative when for any cluster k the within cluster probability of linkage is larger than the probability of linking between clusters k and ℓ , $\forall \ell \neq k$.

For covariate clustering, it is common to make distributional assumptions; usually a mixture model with well-separated centers suffices to show consistency. Dasgupta and Schulman [12] use a Gaussian mixture model, and propose a variant of EM algorithm that provably recovers the center of each Gaussian when the minimum distance between clusters is greater than some multiple of the square root of dimension. Awasthi and Sheffet [3] work with a projection based algorithm and derive conditions on the separation between the cluster centers for obtaining perfect clustering in the limit. Another popular technique is based on SDP relaxations. For example, Peng and Wei [29] propose a SDP relaxation for k-means type clustering. In recent work, Mixon et al. [25] show the effectiveness of SDP relaxation with k-means clustering for subgaussian mixtures, provided the minimum distance between centers is greater than the standard deviation of the sub-gaussian times the number of clusters r .

We also assume that the covariates follow a sub-gaussian distribution, and nodes in the same community share the same mean. In order to match the information content from the network, we consider a covariate model under

two different regimes; namely the low noise and high noise regimes. In the dense graph regime, we use the high dimensional signal plus noise model (low noise) used by [14] extended to sub-gaussian distributions. On the other end of the spectrum, when the graph is sparse, we use a high noise covariate model. Interestingly, similar to [17], in this case we show that, while the error does not go to zero, it is a constant governed by the separation between the means. Our error bound is also similar to [25].

We prove that if the network does not have enough information about separating a cluster out, but the covariates do, or vice versa, it is possible to correctly cluster those nodes with high probability using our convex relaxation. In both regimes, when the separations between a block and the rest are not aligned (orthogonal information being an extreme case of this), we show that combining the two sources can improve the overall accuracy. A byproduct of our theoretical analysis is a simple alternative proof of consistency of SDPs for weakly assortative block models with unequal cluster sizes. Our result also establishes that SDP relaxations can give nearly perfect clustering in weakly assortative models in the sparse regime when cluster sizes are not equal.

In Section 2, we introduce relevant notation and present our optimization framework. We also present the connection of this framework to other well-known methods for community detection both in the dense and sparse regimes. In Section 3, we present our main result. Majority of the proofs are presented in Sections B and C, with details deferred to the appendix. Finally in Section 4, we present experimental results on simulations and real world networks.

2 Problem Setup and Related Work

Assume (C_1, \dots, C_r) represent a r -partition for n nodes $\{1, \dots, n\}$. Let $m_i = |C_i|$ be the size of each cluster, and let m_{\min} and m_{\max} be the minimum and maximum cluster sizes respectively. We will use $\alpha = m_{\max}/m_{\min}$. We denote by A the $n \times n$ binary adjacency matrix and by Y the $n \times d$ matrix of d dimensional covariates. The generation of A and Y share the true and unknown membership matrix $Z = \{0, 1\}^{n \times r}$, for $i \neq j, a, b \in [r]$,

$$P(A_{ij} = 1|Z) = Z_i^T B Z_j \quad \text{For } i \neq j \quad (1)$$

For the graph, B is a $r \times r$ matrix of within and across cluster connection probabilities. Self loops are not allowed. $\max_{ij} B_{ij} = \Theta(1/n)$ in the sparse regime, and $\Omega(\log n/n)$ in the dense regime.

$$Y_i = \sum_{a=1}^r Z_{ia} \mu_a + \frac{W_i}{\sqrt{d}} \quad (2)$$

For the covariates, W_i are mean zero d dimensional subgaussian vectors with spherical covariance matrices $\sigma_k^2 I_d$ and subgaussian norm ψ_k (for $i \in C_k$). We now provide the definition of a subgaussian vector from [32].

Definition 1. The sub-gaussian norm of X is denoted by $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}$. A random vector $X \in \mathbb{R}^n$ is defined to be sub-gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$ with sub-gaussian norm $\|X\|_{\psi_2} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_2}$. Every sub-gaussian random variable X satisfies $P(|X| > t) \leq \exp(1 - ct^2/\|X\|_{\psi_2}^2)$ for all $t \geq 0$.

We consider two regimes for the covariates. In the low noise setting, our model is similar to [14]. We assume that σ_k and ψ_k are both fixed w.r.t n and $d = \Omega(\log n)$. When $\mu_k, k \in [r]$ are sparse, we get low dimensional signal obscured by high dimensional noise. In the low noise case, the signal, i.e. $\|\mu_k - \mu_\ell\|$ for $k, \ell \in [r]$ are fixed w.r.t n ; also the magnitude of the noise is $O_P(1)$. In this case, the kernel matrix concentrates around a suitably defined blockwise constant matrix [14, 35].

In contrast, in the high noise setting, σ_k and ψ_k can depend on d . In particular, when $\sigma_k = \Theta(\sqrt{d})$, our covariate model reduces to a spherical subgaussian mixture model without the \sqrt{d} normalization.

Example In Figure 1(a) we provide an intuitive example of the above setting. Assume that one only observes a $n \times n$ matrix K whose elements are the number of common neighbors between two nodes in a graph, whose adjacency matrix (H) is unobserved. K has been widely used for clustering and link prediction. Now consider the reverse problem of finding a representation Y of the nodes such that $YY^T = K/n$. One factorization is $Y = VE^{1/2}/\sqrt{n}$ where V, E denote the eigenvectors and eigenvalues of K . Since K is positive semi-definite, we can take square root of E . Now we will show that Y_i can be written as $\mu_a + W_i/\sqrt{d}$, for $i \in C_a, a \in \{1, 2\}$. We have $d = n$, and μ_a is a n dimensional sparse vector with signal along the two principal directions. As we will show in the next paragraph, W_i/\sqrt{d} is a vector with norm about $O_P(1)$. While $W_i, i \in [n]$ are not independent, this gives a concrete example of low dimensional signal obscured by high dimensional noise.

Note that $V, E^{1/2}$ also give the eigenvectors and singular values of the unobserved H . If H is generated from a two cluster blockmodel, then for a broad parameter regime, the top two eigenvectors V_{n-1}, V_n concentrate around (rotations of) their population counterparts, which are blockwise constant, with elements $\Theta(1/\sqrt{n})$ [31]. The other eigenvectors are generally believed to be delocalized (shown for centered adjacency matrices in [22]). Hence, for $j < n-1$, $Y_{ij} = O_P(1/\sqrt{n})$, since the bulk eigenvalues are $O_P(\sqrt{n})$; whereas for $j \geq n-1$, Y_{ij} are constant plus $O(1/\sqrt{n})$ noise. In Figure 1(a) we plot the Y matrix and we zoom into the last two columns corresponding to the piecewise constant sparse signal underlying Y . Note that for r blocks the signal will lie along r principal eigenvectors.

2.1 Notation

For a matrix $M \in \mathbb{R}^{n \times n}$, we define the Frobenius and ℓ_∞ norms as: $\|M\|_F^2 = \sum_{i,j} M_{ij}^2$, $\|M\|_\infty = \max_{i,j} |M_{ij}|$. For two matrices $M, Q \in \mathbb{C}^{m \times n}$, their inner product is $\langle M, Q \rangle = \text{trace}(M^T Q)$. The operator norm $\|M\|$ is simply the largest singular value of M , which for a symmetric M , is also the magnitude of its largest eigenvalue. The nuclear norm is $\|M\|_*$ is defined as the sum of the singular values. The $\ell_\infty \rightarrow \ell_1$ norm of a matrix M is defined as $\|M\|_{\ell_\infty \rightarrow \ell_1} = \max_{\|s\|_\infty \leq 1} \|Ms\|_1$. From now on we use I_n to denote the identity matrix of size n , $\mathbf{1}_n$ to represent the all one n -vector and $E_n, E_{n,k}$ to represent the all one matrix with size $n \times n$ and $n \times k$ respectively. We will use $x = \Theta(y)$ to denote that x is some constant times y .

2.2 Existing relaxations for block model inference and kernel clustering

With the exception of [2], most analysis are for strongly assortative block models, which we define below. It is straightforward to see that $E_n - A$'s conditional expectation is assortative, if A comes from a dissortative model. Thus the results are also applicable to strongly dissortative models.

Definition 2. *We will call a stochastic block model strongly assortative if $\min_k B_{kk} > \max_{k \neq \ell} B_{k\ell}$. We will call it weakly assortative if $\forall k \neq \ell, B_{kk} > B_{k\ell}$.*

It is well known [11] that for planted partition models with $B = (p - q)I_r + qE_r$, the log likelihood of the blockmodel can be written as

$$\begin{aligned} & \sum_{i \neq j} \sum_{a,b} Z_{ia} Z_{jb} \left(A_{ij} \log \frac{B_{ab}}{1 - B_{ab}} + \log(1 - B_{ab}) \right) \\ &= \log \frac{p(1-q)}{q(1-p)} \text{trace}(Z^T A Z) + \log \frac{1-p}{1-q} \text{trace}(Z^T E_n Z) + \text{const} \end{aligned}$$

For equal sized clusters, the maximum likelihood problem can be written as an optimization as follows:

$$\hat{Z} = \arg \max_{Z \in \{0,1\}^{n \times r}, Z^T \mathbf{1}_n = \frac{n}{r} \mathbf{1}_r} \text{trace}(Z^T A Z)$$

The above problem is NP-hard due to the intractability of the constrained set. Much work has been done on maximizing the objective under various relaxations. Non-convex approximations naturally lead to the spectral method, where consistency is analyzed by [24, 27, 30] under the dense regime. In the sparse regime, one needs to apply regularization to ensure better classification accuracy than the random predictor [1, 23].

One could also relax the problem to a semidefinite program [2, 9, 11, 17]. SDP's have been shown to be consistent when the graph is dense and all clusters are equal-sized by Amini and Levina [2], Chen and Xu [11]. Furthermore, Guédon and Vershynin [17] show that in the sparse regime one can use this method to obtain an error rate which is a constant w.r.t n that depends on the gap between the within and across cluster probabilities.

Since $ZZ^T \mathbf{1}_n = \frac{n}{r} \mathbf{1}_n$, for equal sized clusters, Amini and Levina [2] use:

$$\begin{aligned} & \max_X \langle A, X \rangle, \\ & \text{s.t. } X \succeq 0, \quad 0 \leq X \leq 1, \quad X \mathbf{1}_n = \frac{n}{r} \mathbf{1}_n, \quad \text{diag}(X) = \mathbf{1}_n \end{aligned} \quad (\text{SDP})$$

For unequal sized clusters, we propose the following SDP:

$$\begin{aligned} & \max_X \langle A, X \rangle, \\ & \text{s.t. } X \succeq 0, \quad 0 \leq X \leq \frac{1}{m_{\min}}, \quad X \mathbf{1}_n = \mathbf{1}_n, \quad \text{trace}(X) = r \end{aligned} \quad (\text{SDP-net})$$

In contrast to **SDP**, **SDP-net** uses a normalized variant of the clustering matrix ZZ^T as X and the desired solution is $X_0 = UU^T$, where $U = Z \text{diag}(\frac{1}{\sqrt{m_1}}, \dots, \frac{1}{\sqrt{m_r}})$;

$\|X_0\|_F^2 = r$. A similar formulation has been used by [25]. The m_{\min} in the constraint can be replaced by any lower bound on the smallest cluster size.

Most results focus on the strongly assortative case except Amini and Levina [2], who show that in the dense regime **SDP** is consistent even for weakly assortative block models with equal sized clusters. Our result however only establishes weak consistency. The authors also propose a semidefinite relaxation for unequal sized clusters (discussed in Section 4) which is shown to yield reasonable clustering for the population matrix $E[A|Z]$. In Section 4, we show that the constraint set of this SDP requires much heavier computation than **SDP-net**.

For covariate clustering, for ease of exposition, we start with the simple case of multivariate Gaussian covariates with the same spherical covariance, i.e., $Y_i \sim \mathcal{N}(\mu_k, \sigma^2 I)$ if $i \in C_k$. We denote $\boldsymbol{\mu} \in \mathbb{R}^{r \times d} = (\mu_1, \dots, \mu_r)^T$. Now the log-likelihood of Y equals the following up-to additive constants:

$$-\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - Z_i \boldsymbol{\mu})^T (Y_i - Z_i \boldsymbol{\mu}) = -\frac{1}{2\sigma^2} \text{trace}((Y - Z\boldsymbol{\mu})^T (Y - Z\boldsymbol{\mu})) \quad (3)$$

If all clusters have equal sizes, the MLE of $\boldsymbol{\mu}$ for known Z is none other than the empirical centroid $\hat{\boldsymbol{\mu}} = (\text{diag}(Z^T \mathbf{1}_n))^{-1} Z^T Y = Z^T Y / \frac{n}{r}$. Substituting $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$ in Equation 3 gives:

$$-\frac{1}{2\sigma^2} \text{trace}((Y - Z\hat{\boldsymbol{\mu}})^T (Y - Z\hat{\boldsymbol{\mu}})) = \frac{r}{2n\sigma^2} \text{trace}(YY^T ZZ^T) + \text{const}$$

Thus maximizing the above function in the case of mixture of Gaussian with equal sized components with identical spherical variance boils down to maximizing $\text{trace}(YY^T X)$ over a semidefinite matrix X . As it turns out, a SDP relaxation of the k-means loss also has the exact same form [28].

$$\sum_{k=1}^r \sum_{i: Z_{ik}=1} \|Y_i - \hat{\boldsymbol{\mu}}_k\|^2 = \frac{r}{n} \text{trace}(YY^T ZZ^T) + \text{const}$$

The linear inner product can be further generalized to Gaussian kernel (defined in Section 3) to capture the non-linearity in the data space. Ignoring all the constants the objective is now $\text{trace}(KX)$ for some kernel similarity matrix K which we formally define later. Several convex relaxations for k-means type loss are proposed in the literature (see [25] for more references). It has also been shown [13] that kernel spectral clustering is equivalent to weighted kernel k-means by maximizing $\text{trace}(Z^T K Z)$.

2.3 Existing work incorporating covariates

In most real world examples, covariates on nodes of the network are available and they reflect the latent clustering structure in a certain way. Some recent works have been focussed on incorporating covariates for community detection [5, 18, 34, 38]. In [5], the authors present assortative covariate-assisted spectral clustering (ACASC) where one adds YY^T to the regularized graph Laplacian weighted by a tuning parameter. The clusters are estimated using Spectral Clustering on top eigenvectors of the sum.

A joint criterion for community detection (JCDC) with covariates is proposed by [38], which could be seen as a covariate reweighted Newman-Girvan

modularity. This approach enables learning different influence on each covariate. The assumptions include strong assortativity and bounded covariates. In their work, the optimization is carried out in an alternative manner, which is a jointly non-convex problem. Thus although the true maximizer of the JCDC objective is consistent, the algorithm may return a local optima. Notable Bayesian and likelihood-based approaches that include covariates for block models appear in [26, 34, 37].

2.4 Our contribution

In this paper, we propose to add a k-means type penalty term to the objective which enforces that the estimated clusters are consistent with the latent memberships in the covariate space. In particular, we use a kernel to allow for non-linear boundaries between clusters. Let K be the $n \times n$ kernel matrix whose (i, j) -th entry is $K(y_i, y_j)$. Our SDP relaxation is as follows:

$$\begin{aligned} & \max_X \quad \langle A + \lambda K, X \rangle, \\ \text{s.t. } & X \succeq 0, \quad 0 \leq X \leq \frac{1}{m_{\min}}, \quad X \mathbf{1}_n = \mathbf{1}_n, \quad \text{trace}(X) = r \end{aligned} \quad (\text{SDP-comb})$$

Here λ is a tuning parameter. If one uses $K(y_i, y_j)$ as the dot product $y_i^T y_j$, then the non-convex variant of the objective function naturally assumes a form similar to the work of ACASC (modulo normalization of A).

Although the analysis can be generalized to other kernel functions, for simplicity we focus our analysis on the Gaussian kernel. This kernel function is upper bounded by 1 and is Lipschitz continuous w.r.t. the distance between two observations. In addition to showing that our SDP relaxation is consistent in the dense case and achieves a better error bound in the sparse case, we show that if the information from the graph and the covariates is complementary, i.e. there exist clusters that the network cannot separate out, whereas the covariates can (or vice versa), our method is able to successfully achieve (almost) perfect clustering for the entire network in the (sparse) dense regime in the limit.

Our contribution thus is two-fold. In order to operate in a setting where both sources can be equally powerful, we analyze two regimes; sparse graphs with high noise covariates and dense graphs with low noise covariates. We show that when the information from the two sources is not aligned, combining them improves the error bound over a broader range of parameters. In addition, we obtain a similar result as in [17] for sparse graphs, which also works for weakly assortative models.

3 Main result

It has been well established that sparse and dense graphs have very different theoretical properties in terms of concentration [23]. Analogously, kernel matrices have different properties in the low noise and high noise settings. To be more specific, when the graph is dense, the adjacency matrix concentrates to its expectation in operator norm [27]. Similarly, in the low noise case, the kernel matrix concentrates elementwise and in Frobenius norm around a suitably defined limit [14, 36], which is blockwise constant in our subgaussian mixture

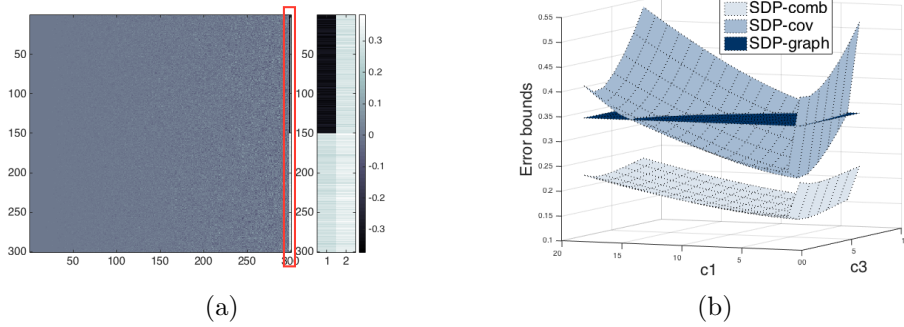


Figure 1: (a) Example of low dimensional signal plus high dimensional noise. Every row in this matrix is a datapoint in 300 dimensional space. The zoom in shows the last two columns corresponding to the underlying sparse signal. (b) Error surfaces for sparse graph, high noise covariates and their combination.

setting. In contrast to the dense case, when graph is sparse, the SDP is not consistent, but can achieve an error rate governed by the within and across cluster probabilities. We present an analogous result for high noise covariates, which establishes that, while SDP with high noise covariates is not consistent, it achieves a small error rate if the cluster centers are further apart.

We first present the result on dense graph and low noise covariates in subsection 3.1, followed by the result for sparse graph and high noise covariates in subsection 3.2. Throughout the remainder of this manuscript, K is the kernel matrix obtained from data, $K(i, j) = f(\|Y_i - Y_j\|_2^2)$, where $f(\cdot)$ is the gaussian kernel function. To be concrete, $f(x) = \exp(-\eta x)$.

Let $p_k := B_{kk}$, $q_k := \max_{\ell \neq k} B_{k\ell}$, $p_{\max} = \max_k p_k$ and $p_{\min} = \min_k p_k$. The distance between clusters C_k and C_ℓ is

$$d_{k\ell} = \|\mu_k - \mu_\ell\|, \quad d_{\min} = \min_{k \neq \ell} d_{k\ell}$$

3.1 Dense graph and low noise covariates

In this section we assume that each covariate consists of low-dimensional information plus high dimensional noise (Equation 2). This model allows the distance between the cluster means to remain constant as n grows. Since K concentrates around a blockwise constant matrix, we define the separation for K as:

$$\nu_k := f(2\sigma_k^2) - \max_{\ell \neq k} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2), \quad (4)$$

We will use ν_k as the separation corresponding to the covariate clusters; however it is defined differently in the low noise and high noise cases.

Theorem 1 (Dense graph plus low noise covariates). *Define ν_k as in Equation 5. For $\gamma' = \min_k \left(\frac{p_k - q_k}{1 + \lambda} + \frac{\lambda}{1 + \lambda} \nu_k \right) \geq 0$, we have:*

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{\sqrt{2\alpha^2 r}}{\gamma'} \left(\frac{1}{1 + \sqrt{\lambda}} C_G \sqrt{\frac{r p_{\max}}{n}} + \frac{\sqrt{\lambda}}{1 + \sqrt{\lambda}} C_K \sqrt{\frac{\log n}{\min(d, n)}} \right)$$

REMARK ON ORTHOGONAL INFORMATION CONTENT: Compared to prior work, the separation condition in Theorems 6 and 1 indicates that adding the kernel part allows the SDP to succeed in a larger range of situations. In particular, for each cluster, both of $p_k - q_k$ and ν_k do not have to be large positive numbers as long as their sum is. We want to point out that, if one of the sources has much stronger signal, it may render the other source useless asymptotically. However, if $p_k - q_k$ and $\lambda\nu_k$ are of the same order and the separation conditions are satisfied, then one would get a constant times better improvement in the error rate.

The most interesting scenario arises however, where for some k , $p_k - q_k$ is not large enough to satisfy the separation condition on the network, however ν_k is large and hence putting the two together has a better chance at separating out the k^{th} cluster. That said, it is also possible to get a worse clustering, if one was adversarial and one source was forced to have enough noise to drown the signal from the other. One can, in principle, get around this problem by tuning λ . The tuning procedure is discussed in detail in Section 4.

The proof of our main result relies on two intermediate results which can be also seen as two special cases of **SDP-comb**. Setting $\lambda = 0$ we have the following result for weakly assortative block models in the dense case. On the other extreme, by taking $\lambda \rightarrow \infty$ in **SDP-comb** we provide conditions for the consistency of kernel clustering for low noise covariates.

Theorem 2 (Analysis of dense graph). *Let \hat{X} be the optimal solution of **SDP-***

net. If $\min_k(p_k - q_k) > 0$, then, with probability tending to one, $\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{r\alpha}{\min_k(p_k - q_k)} O_P\left(\sqrt{\frac{p_{\max}}{n}}\right)$. Therefore,

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} = o_P(1) \quad \text{if} \quad \min_k(p_k - q_k)/r\alpha = \Omega(\sqrt{p_{\max}/n})$$

REMARK ON STRONG VS. WEAK CONSISTENCY: A strong consistency result ($\hat{X} = X_0$ w.h.p) for a slightly different SDP was shown in the dense case by [2], under the assumption of equal cluster sizes. In fact, our proof technique can be adapted to recover their result under the same condition. However, under the more general setting of unequal clusters, we can only show weak consistency, i.e. the suitably normalized Frobenius norm of the error converges to zero.

We now present the result in the low noise covariate setting.

Theorem 3 (Low noise covariates). *Let \hat{X} be the solution of the following SDP,*

$$\begin{aligned} \max_X \quad & \langle K, X \rangle, \\ \text{s.t.} \quad & X \succeq 0, \quad 0 \leq X \leq 1/m_{\min}, \quad X\mathbf{1}_n = \mathbf{1}_n, \quad \text{trace}(X) = r. \end{aligned} \quad (\text{SDP-cov})$$

where K is the kernel matrix generated from kernel function f . Denote

$$\nu_k := f(2\sigma_k^2) - \max_{\ell \neq k} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2), \quad (5)$$

and $\alpha := m_{\max}/m_{\min}$. If $\min_k \nu_k \geq 0$, for $p = \Omega(\log n)$, there exists a constant C_K , such that with probability tending to one, we have:

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{C_K}{\min_k \nu_k} \sqrt{\frac{r\alpha^2 \log n}{\min(d, n)}}$$

We have $\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} = o_P(1)$, if $\min_k \nu_k = \Omega\left(\sqrt{\frac{r\alpha^2 \log n}{\min(d, n)}}\right)$.

Since we use the Gaussian kernel, ν_k is bounded above by one. While the exact value of ν_k depends on the choice of η and the model parameters, lower variance and larger inter-cluster distance will result in larger ν_k . It is important to note that under our model assumption, $d_{k\ell}$ does not increase as dimension d goes to infinity.

3.2 Sparse graph and high noise covariates

Since the analysis for sparse graph and high noise covariates is different from the last setting, we will first provide our result on sparse graphs, which is similar to that of [17]; in contrast to their work, it is also applicable to the weakly assortative case.

Theorem 4 (Analysis for sparse graph). *Let \hat{X} be the solution to **SDP-net**, $p_k = a_k/n, q_k = b_k/n$, where a_k, b_k are positive constants. Assume that $\bar{p} := \frac{2}{n(n-1)} \sum_{i < j} \text{Var}(a_{ij}) =: \frac{g}{n}$. If $g \geq 9$, then with probability tending to 1,*

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \epsilon,$$

if $\min_k (a_k - b_k) \geq \frac{23\alpha^2 r \sqrt{g}}{2}$ where $\alpha := m_{\max}/m_{\min}$.

Note that in the above theorem, in order to have the error rate ϵ to go to zero, one would require $a_k - b_k$ to go to infinity, whereas by definition a_k, b_k are constants. Therefore one can only hope for a small albeit constant ϵ . We would like to point out that, in contrast to having $\min_k a_k - \max_k b_k$ (strong assortativity) in the denominator like [17], we have $\min_k (a_k - b_k)$ (weak assortativity), which allows for a much broader parameter regime.

REMARK ON THE EFFECT OF UNEQUAL SIZES: Note that in the sparse case, we need α and r to be $\Theta(1)$ in order to satisfy the separation condition for $\min_k (p_k - q_k)$. In the dense case, (Theorem 2) $r\alpha = O(\sqrt{n})$ for satisfying the separability condition on $\min_k (p_k - q_k)$. Which means, for balanced clusters $r = O(\sqrt{n})$.

Now we present an analogous result for high noise covariates. This will motivate the final theorem on $A + \lambda K$. The main difficulty is that like the low noise case, one cannot achieve element-wise or operator norm concentration of the kernel matrix in the high noise case (also discussed in [33]). Hence, we divide the nodes into “good” nodes $\mathcal{S}_k := \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}$. Also define $\mathcal{S} = \cup_{k=1}^r \mathcal{S}_k$. Let

$$r_k := f(2\Delta_k), \quad s_k := \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell) \quad (6)$$

This is true because, a simple use of triangle inequality gives $\min_{i,j \in \mathcal{S}_k} K_{ij} \geq r_k$ and $\max_{i \in \mathcal{S}_k, j \in \mathcal{S}_\ell, \ell \neq k} K_{ij} \leq s_k$. We define the separation for cluster k as:

$$\nu_k = r_k - s_k \quad (7)$$

Δ_k will be defined such that the kernel matrix induced by the rows and columns in \mathcal{S} is weakly assortative, i.e. $\min_k \nu_k \geq 0$. To make the analysis concrete, for Theorem 5, we use $\Delta_k = \Delta$. The choice of Δ is crucial. A large Δ makes the size of non-separable nodes \mathcal{S}^c small, but drives down the separation ν_k .

Theorem 5 (High noise covariates). *Let \hat{X} be the optimal solution of **SDP-cov**, $\alpha = \frac{m_{\max}}{m_{\min}}$. If $\frac{d_{\min}}{\psi_{\max}} > \max\{1, \frac{180}{d}\}$, then for a properly chosen η , with probability at least $1 - \sum_k \frac{1}{m_k}$,*

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 \max\left(\frac{\psi_{\max}^2}{d_{\min}^2} \max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}}\right), r\right\}, r\sqrt{\frac{\log m_{\max}}{m_{\max}}}\right)$$

If ψ_{\max} grows as $\theta\sqrt{d}$, when α, r are fixed w.r.t. n , d_{\min}^2 has to be much larger than $\theta^2 d$ in order to drive the error rate down.

Now consider the problem where the covariates are combined with the network. As we will show in the proof, the new separation is $\gamma = \min_k((p_k - q_k) + \lambda\nu_k)$. Typically, with unequal subgaussian norms, one should benefit from using different Δ_k 's for different clusters. For example for a cluster with a large $p_k - q_k$, we can afford to have a small ν_k . To think in terms of Δ_k , for this cluster one can have a large Δ_k , which will make $|\mathcal{S}_k|$ larger than before, but will not affect the separation of cluster k $(p_k - q_k) + \lambda\nu_k$ very detrimentally. We now present the theorem combining the two sources of information.

Theorem 6 (Sparse graph plus high noise kernel). *Let $p_k = a_k/n, q_k = b_k/n, g = \bar{p}/n$. Using $\lambda = \ell/n, m_k = n\pi_k, m_{\min} = n\pi_{\min}$, and $\pi_0 := \sum_k (m_k \exp(-c_k^2 d) + \sqrt{m_k \log m_k/2})/n$, we get:*

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq 4K_G \frac{6\sqrt{g} + \ell(2\pi_0 + \sum_k \pi_k^2(1 - f(2\Delta_k)))}{r\pi_{\min}^2 \min_k(a_k - b_k + \ell\nu_k)},$$

where $\nu_k = f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$ for some Δ_k such that $\min_k(a_k - b_k + \ell\nu_k) \geq 0$.

If covariates have much stronger signal than the graph, ℓ should be large. However, when both sources have equally strong signals, we take ℓ as a constant. In general the upper bound depends on several parameters such as λ and the scale parameter η in the gaussian kernel. We provide procedures for tuning λ and η in Section 4. The Δ_k 's show up in the numerator as well as the denominator. Finding the optimal Δ_k is cumbersome in the general case with unequal ψ_k 's, and we use a concrete example to show how the combined upper bound can be better than one source.

Consider the following example. Let $d = 2, \alpha = 1, \mu_1 = (400, 0), \mu_2 = (-400, 0), \mu_3 = (0, 400\sqrt{3})$, so that the distances between any two centers are equal. We generate three gaussians with the above center and standard deviation $\psi_1 = \psi_2 = 4, \psi_3 = 50$. Therefore cluster 3 is hard to separate out using the covariates.

For the graph, we use $B = \text{diag}(p_1 - q, p_2 - q, p_3 - q) + qE_3$. We set $p_1 = p_2 = 1/300, p_3 = 1/150$ and $q = 1/30000$. So cluster 3 is easily separable in the network. We now plot three error bounds, one for the graph, one for the covariate and one for the combined SDP. Denoting $\Delta_k := c_k \psi_k \sqrt{d}$, we vary λ, c_1, c_2 and c_3 . Figure 1(b) is shown for $\lambda = .002$, and $c_1 = c_2$. The surface of the error bound in Theorem 6 is shown as c_1 and c_3 are varied. For the error rate of covariates, we take $\ell \rightarrow \infty$ in the bound in Theorem 6.

The network has constant error, since it does not depend on c_1, c_2 or c_3 . Note that by optimizing over c_1, c_2, c_3 one can bring down the error rate of the

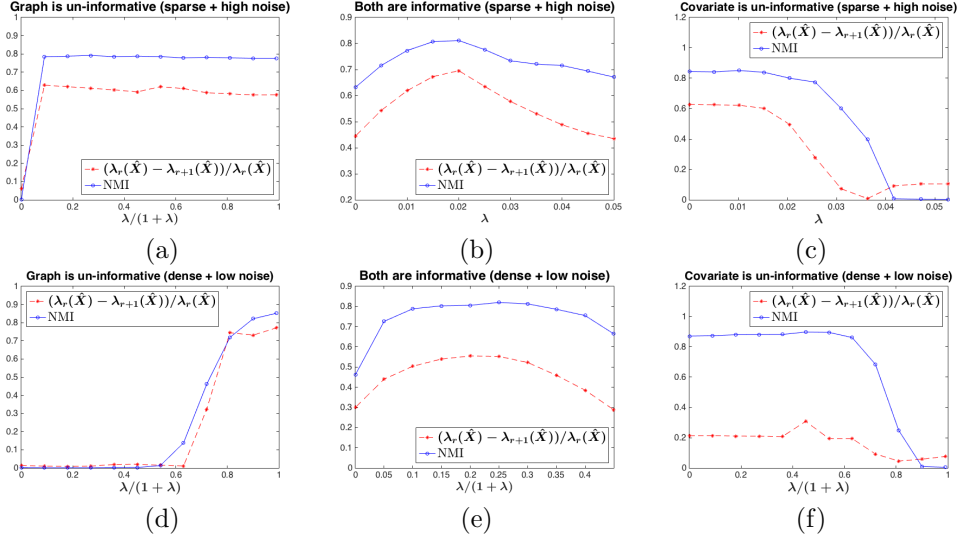


Figure 2: Tuning procedure: (a)-(c) sparse graph with low dimensional covariates, (d)-(f) dense graph with high dimensional covariates. (a) $B = 0.005E_3$, $n = 1000$, $d = 6$, $d_{\min} = 15\sigma$; (b) $d = 6$, $d_{\min} = 1.3$, $\sigma = (1, 1, 5)$, $B = \text{diag}(0.004, 0.024, 0.024) + 0.004E_3$; (c) $d = 6$, $B = 0.0144I_3 + 0.0016E_3$; (d) $B = 0.1E_3$, $n = 1000$, $r = 3$, $d = 50$, $m = (m_0, m_0, 2m_0)$, $d_{\min}^2 = 0.2$, $\sigma = 1$. (e) $d = 100$, $d_{\min}^2 = 0.005$, $\sigma = (1, 1, 2)$, $B = 0.005E_3 + \text{diag}(0.005, 0.02, 0.02)$, (f) $d = 62$, $B = 0.08I_3 + 0.04E_3$.

covariates. However, by adding in the network, one can achieve a better error than either source as shown in Figure 1(b).

We now present the tuning procedure, and experimental results.

4 Experiments

In this section, we present results on real and simulated data. Additional experimental results are deferred to the appendix. Performance of the clustering is measured by normalized mutual information (NMI), which is defined as the mutual information of the two distributions divided by square root of the product of their entropies. We have also calculated classification accuracy and they show similar trends. Our implementation of SDP uses the alternating descent method of multipliers (ADMM) algorithm [6]. The implementation details can be found in the appendix. We start with a discussion of the method of tuning.

4.1 Choice of tuning parameters

SDP-comb has three tuning parameters, η , λ , and m_{\min} . In order to pick η , we do a grid search from $\eta \in [10^{-5}, 10^5]$ and calculate \hat{X}_η using **SDP-cov**. We pick the η which maximizes $(\lambda_r(\hat{X}_\eta) - \lambda_{r+1}(\hat{X}_\eta))/\lambda_r(\hat{X}_\eta)$. Let $g(\cdot) = (\lambda_r(\cdot) - \lambda_{r+1}(\cdot))/\lambda_r(\cdot)$.

As λ increases, the resulting **SDP-comb** clustering gradually changes from the **SDP-net** clustering to the **SDP-cov** clustering. Our theoretical consistency

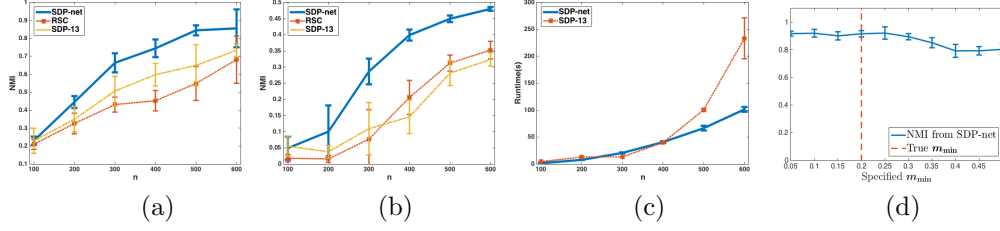


Figure 3: $r = 3$: (a) Strongly assortative graph; $m = (m_0, 2m_0, 2m_0)$, $B = \begin{bmatrix} 0.42 & 0.20 & 0.32 \\ 0.20 & 0.38 & 0.28 \\ 0.32 & 0.28 & 0.42 \end{bmatrix}$; (b) Weakly assortative: $m = (m_0, m_0, 3m_0)$, $B = \begin{bmatrix} 0.2139 & 0.0689 & 0.1888 \\ 0.0689 & 0.1041 & 0.0790 \\ 0.1888 & 0.0790 & 0.2139 \end{bmatrix}$; (c) Runtime for the setting in (a); (d) Robustness to the choice of m_{\min} , $n = 500$, $m = (m_0, 2m_0, 2m_0)$, $B = 0.1E_3 + 0.1I_3$.

results show that, with the right λ , \hat{X}_λ and X_0 should be close, and hence also have similar eigenvalues. Using Weyl's inequality and the fact that $\|\hat{X} - X_0\|_{\text{op}} \leq \|\hat{X} - X_0\|_F$,

$$\lambda_r(X_0) - \|\hat{X} - X_0\|_F \leq \lambda_r(\hat{X}) \leq \lambda_r(X_0) + \|\hat{X} - X_0\|_F$$

While we do not know the eigenvalues of X_0 , we use the fact that $g(X_0) = 1$. Hence we pick the λ maximizing $g(\hat{X}_\lambda)$. Figure 2 (a)-(c) are for sparse graph and low dimensional covariates, and (d)-(f) are for dense graph and high dimensional covariates. In each regime, figures from left to right represent the situation where graph is uninformative (Erdős-Rényi), both are informative and covariates are uninformative. We plot $g(\hat{X}_\lambda)$ and NMI of the clustering from \hat{X}_λ with the true labels against λ . For high dimensional regime we search over $\lambda/(1+\lambda) \in [0, 1]$, whereas for the sparse and low dimensional covariates regime, $\lambda = \Theta(1/n)$. Figure 2 shows that $g(\hat{X}_\lambda)$ and NMI of the predicted clustering have a similar trend, justifying the effectiveness of the tuning procedure.

Finally, we need to tune m_{\min} . Recall that any lower bound of m_{\min} ensures consistent clustering from **SDP-comb**. In all the experiments below, we use m_{\min} as the true minimum of the cluster sizes. In Figure 3-(d) we show that **SDP-comb** is robust to a broad range of m_{\min} values.

4.2 Comparison with alternative SDP's and Regularized Spectral Clustering

Now we compare **SDP-net** with the proposal of [2] for weakly and strongly assortative cases, in terms of NMI and runtime, and show the robustness of **SDP-net** to the choice of m_{\min} . We generate networks from stochastic block models and compare the clustering from **SDP-net** with state-of-the-art methods: 1) regularized spectral clustering (RSC) in [1] and 2) SDP-13 (for imbalanced clusters) in [2]. Let \mathcal{F}_{13} be the set $\{X : X \succeq 0, X \geq 0, X\mathbf{1}_n \geq m_{\min}\mathbf{1}_n, \text{diag}(X) = \mathbf{1}_n\}$.

$$\max_X \langle A - \mu E, X \rangle, \quad s.t. \quad X \in \mathcal{F}_{13} \quad (\text{SDP-13})$$

In each case the average NMI from 10 randomly generated networks is reported. In Figures 3(a) and (b) we plot NMI on the Y axis vs n on the X

axis; (a) and (b) respectively show results for a strongly and weakly assortative model. It can be seen that the NMI for all methods increase as n grows, but **SDP-net** outperforms the other methods. Figure 3(c) shows that as the size of graph grows, **SDP-comb** runs significantly faster than **SDP-13**. This is due to the structure of the feasible set of **SDP-comb**, which leads to simple and computationally efficient projections onto the linear space. In contrast, the ADMM for solving **SDP-13** needs n projections to half spaces in each iteration. Finally, Figure 3(d) shows that the recovery is fairly robust within a rather broad range of mis-specified m_{\min} .

4.3 Real World Networks

Now we present results on a real world social network and a ecological network. We compare: (1) SDP on network only [2]; (2) kernel clustering with covariates (**SDP-cov**); (3) Covariate-assisted spectral clustering (ACASC) [5]; (4) JCDC [38]; and (5) our **SDP-comb**.

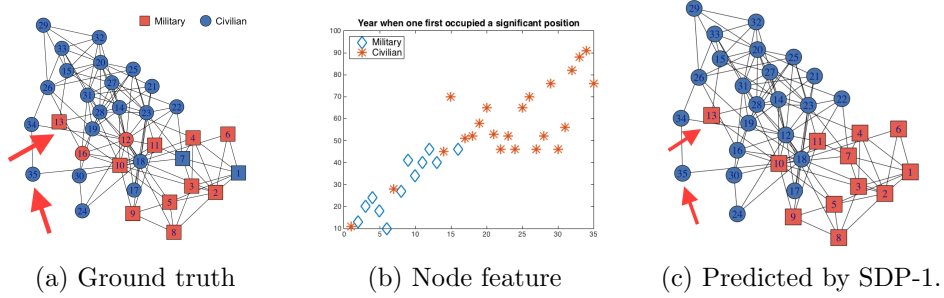


Figure 4: Mexican political network: The nodes which have equal connection to both groups, can be distinguished by the year in which they came into power. The covariates which are close to the time of transition may be distinguished by their connections in the network.

Mexican political elites This network [15] depicts the political, kinship, friendship, or business interactions between 35 Mexican political elites, e.g. presidents and close collaborators, etc. The political arena consists of two main groups, namely the military and the civilians. The earlier part of the twentieth century was dominated by the military who came to power after the revolution. The military were succeeded by the civilians. We use the professional background (military or civilian) as the ground truth label. The year in which a politician first held a significant governmental position, is used as a covariate. Figure 4(a) shows that the covariate gives a good indication of the labels.

We compare **SDP-net**, **SDP-comb**, **SDP-cov**, JCDC and ACASC. In Figure 4(d-f), we can see the improvement from leveraging both covariates and graph. For example, both node 35 and 13 have exactly one connection with the military and one with the civilian group. By taking the covariate into consideration, we can correctly assign the labels to both of these ambiguous nodes. Table 1 shows the NMI of all methods, where our method outperforms other covariate-assisted approaches.

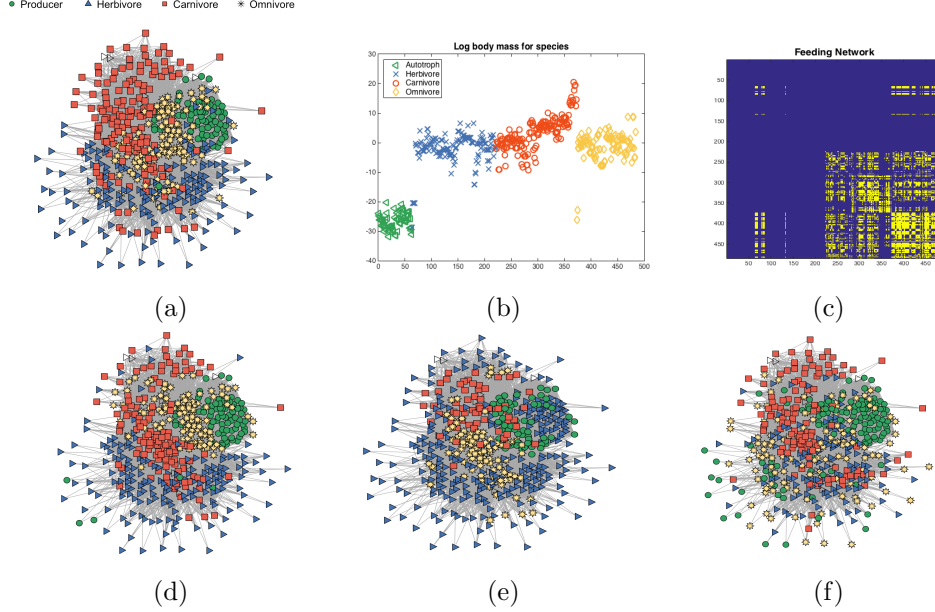


Figure 5: Weddell sea network: (a) Ground truth labels; (b) Log body mass by species; (c) Constructed adjacency matrix A_τ ; (d) Labeled by **SDP-comb**; (e) Labeled by **SDP-net**; (f) Labeled by **SDP-cov**.

Weddell sea trophic dataset The next example we consider is a ecological network collected by [21] describing the marine ecosystem of Weddell Sea, a large bay off the coast of Antarctica. The dataset lists 489 marine species and their directed predator-prey interactions, as well as the average adult body mass for each of the species. We use a thresholded symmetrization of the directed graph as the adjacency matrix. Let G be the directed graph, the $(i, j)^{th}$ entry of GG^T captures the number of other species which i and j both feed on. We create binary matrices $A_\tau = 1(GG^T \geq \tau)$. Choosing different τ 's between 1 to 10 gives similar clustering. We use $\tau = 5$.

All species are labeled into four categories based on their prey types. Autotrophs (e.g. plants) do not feed on anything. Herbivores feed on autotrophs. Carnivores feed on animals that are not autotrophs, and the remaining are omnivores, which feed both on autotrophs and other animals (herbivore, carnivore, or omnivores). Since body masses of species vary largely from nanograms to tons, we work with the normalized logarithm of mass following the convention in Newman and Clauset [26]. Figure 5(b) illustrates the log body mass for species. Without loss of generality, we order the nodes as autotrophs, herbivores, carnivores and omnivores.

In Figures 5(c), we plot A_τ . Since the autotrophs do not feed on other species in this dataset, and since herbivores do not have too much overlap in the autotrophs they feed on, the upper left corner of the input network is extremely sparse. On the other side, the body sizes for autotrophs are much smaller than those of other prey types. Therefore the kernel matrix clearly separates them out.

We see that **SDP-net** (Figure 5(e)) heavily misclusters the autotrophs since it

Dataset	SDP-net	SDP-cov	SDP-comb	ACASC	JCDC
Mexican politicians	0.37	0.43	0.46	0.37	0.25
Weddell Sea	0.36	0.22	0.50	0.32	0.42

Table 1: NMI with ground truth for various methods

only replies on the network. **SDP-cov** (Figure 5(f)) only takes the covariates into account and cannot distinguish herbivores from omnivores, since they possess similar body masses. However, **SDP-comb** (Figure 5(d)) achieves a significantly better NMI by combining both sources. Table 1 shows the NMI between predicted labels and the ground truth from **SDP-comb**, JCDC and ACASC. While JCDC and ACASC can only get as good as the the best of graph or covariates, our method achieves a higher NMI.

In the next two sections, we present the proofs of dense graph plus low noise covariates and sparse graph plus high noise covariates respectively. For some long proofs, we only present a sketch and defer the details to the appendix.

5 Analysis for dense graph and high dimensional covariates

5.1 Analysis of dense graph

We will first present a few results that would be used to prove the main theorems.

Lemma 1. *For any X that satisfies $X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}$, we have $\|X\|_F^2 \leq \text{trace}(X)$.*

Lemma 2. *If $x^2 - ax - b^2 \leq 0$, for some $a, b > 0$, then $x \leq a + b$.*

Lemma 3. *Let \hat{X} be the optimal solution of **SDP-net**, Q is a matrix where $Q_{ij} = p_k, \forall i, j \in C_k$, and $q_k \geq Q_{ij} \geq 0, \forall i \in C_k, j \in C_\ell, k \neq \ell$. If $\min_k(p_k - q_k) \geq 0$, then*

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min} \min_k(p_k - q_k)} \langle A - Q, \hat{X} - X_0 \rangle$$

Proof of Lemma 3. Note that both X_0 and \hat{X} are in the feasible set of **SDP-net**, by optimality, we have $\langle A, \hat{X} \rangle \geq \langle A, X_0 \rangle$. We construct Q as stated in the lemma to obtain: $\langle Q, \hat{X} - X_0 \rangle$,

$$\langle A - Q, \hat{X} - X_0 \rangle \geq \langle Q, X_0 - \hat{X} \rangle$$

Note that Q is constant on diagonal blocks and upper bounded by q_k on off-diagonal blocks, with respect to the clustering of nodes. Using the fact that

$|C_k| = m_k$, we have:

$$\begin{aligned}
\langle Q, X_0 - \hat{X} \rangle &= \sum_k \sum_{i \in C_k} \left(p_k \sum_{j \in C_k} \left(\frac{1}{m_k} - \hat{X}_{ij} \right) + \sum_{\ell \neq k} \sum_{j \in C_\ell} Q_{ij} (0 - \hat{X}_{ij}) \right) \\
&\geq \sum_k \sum_{i \in C_k} \left(p_k \sum_{j \in C_k} \left(\frac{1}{m_k} - \hat{X}_{ij} \right) - q_k \sum_{\ell \neq k} \sum_{j \in C_\ell} \hat{X}_{ij} \right) \\
&= \sum_k \sum_{i \in C_k} \left(p_k \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right) - q_k \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right) \right) \\
&= \sum_k \sum_{i \in C_k} (p_k - q_k) \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right) \geq \min_k (p_k - q_k) \sum_k \sum_{i \in C_k} \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right)
\end{aligned}$$

The third line and last inequality uses the constraint that $\sum_j \hat{X}_{ij} = 1$, and $1 - \sum_{j \in C_k} \hat{X}_{ij} \geq 1 - \sum_j \hat{X}_{ij} = 0$. On the other hand,

$$\|\hat{X} - X_0\|_F^2 = \|\hat{X}\|_F^2 - \|X_0\|_F^2 + 2\langle X_0 - \hat{X}, X_0 \rangle$$

By Lemma 7, and the fact that $\|X_0\|_F^2 = r$ by construction, we have $\|\hat{X}\|_F^2 - \|X_0\|_F^2 \leq \text{trace}(\hat{X}) - r = 0$. Since $\min_k (p_k - q_k) \geq 0$,

$$\begin{aligned}
\|\hat{X} - X_0\|_F^2 &\leq 2\langle X_0 - \hat{X}, X_0 \rangle = 2 \sum_k \sum_{i \in C_k} \sum_{j \in C_k} \frac{1}{m_k} \left(\frac{1}{m_k} - \hat{X}_{ij} \right) \\
&= 2 \sum_k \sum_{i \in C_k} \frac{1}{m_k} \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right) \leq \frac{2}{m_{\min}} \sum_k \sum_{i \in C_k} \left(1 - \sum_{j \in C_k} \hat{X}_{ij} \right) \\
&\leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \langle Q, X_0 - \hat{X} \rangle \\
&\leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \langle A - Q, \hat{X} - X_0 \rangle
\end{aligned}$$

□

Proof of Theorem 2. Let $Q = ZBZ^T$, then by Lemma 3,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \langle A - Q, \hat{X} - X_0 \rangle \quad (8)$$

Notice that A and $P := E[A|Z]$ has zero diagonals. Therefore,

$$\begin{aligned}
\langle P - Q, \hat{X} - X_0 \rangle &= \sum_k \sum_{i \in C_k} p_k \left(\frac{1}{m_k} - \hat{X}_{ii} \right) \\
&\leq \sum_k p_k - p_{\min} \text{trace}(\hat{X}) \leq r(p_{\max} - p_{\min})
\end{aligned} \quad (9)$$

Recall that U is a $n \times r$ matrix with columns as the eigenvectors of X_0 . Now for any matrix X , define the projection $P_{T^\perp}(X) = (I - UU^T)X(I - UU^T)$, and $P_T(X) = X - P_{T^\perp}(X)$.

$$\langle A - P, \hat{X} - X_0 \rangle = \underbrace{\langle P_{T^\perp}(A - P), \hat{X} - X_0 \rangle}_{R_1} + \underbrace{\langle P_T(A - P), \hat{X} - X_0 \rangle}_{R_2}$$

We will now bound R_1 and R_2 . It is known (see [8]) that the sub-gradient of the nuclear norm of X_0 is

$$\partial\|X_0\|_* = \{UU^T + W, U^TW = WU = 0, \|W\|_2 \leq 1\}$$

Let $W = P_{T^\perp}\left(\frac{A-P}{\|A-P\|_2}\right)$, we have $\|P_{T^\perp}(W)\| \leq \|\frac{A-P}{\|A-P\|_2}\| = 1$; so $UU^T + W$ is in $\partial\|X_0\|_*$. Using the definition of a sub-gradient, we have

$$\langle UU^T + P_{T^\perp}(W), \hat{X} - X_0 \rangle \leq \|\hat{X}\|_* - \|X_0\|_* = 0$$

Since, $X_0 = UU^T$,

$$\begin{aligned} R_1 &:= \langle P_{T^\perp}(A - P), \hat{X} - X_0 \rangle \leq \|A - P\|_2 \langle UU^T, X_0 - \hat{X} \rangle \\ &\leq \|A - P\|_2 \|X_0\|_F \|X_0 - \hat{X}\|_F \leq \sqrt{r} \|A - P\|_2 \|X_0 - \hat{X}\|_F \end{aligned} \quad (10)$$

On the other hand, since $P_T(A - P)$ is of rank at most $3r$, $\|P_T(A - P)\|_F \leq \sqrt{3r} \|P_T(A - P)\|_2$. Now we can bound $\|P_T(A - P)\|_F$ as follows:

$$\|P_T(A - P)\|_F \leq \sqrt{3r} (\|P_{T^\perp}(A - P)\|_2 + \|A - P\|_2) \leq 4\sqrt{r} \|A - P\|_2$$

Therefore:

$$R_2 := \langle P_T(A - P), \hat{X} - X_0 \rangle \leq 4\sqrt{r} \|A - P\|_2 \|X_0 - \hat{X}\|_F \quad (11)$$

Substituting Equations 9, 10, 11 back to 8 we have,

$$\begin{aligned} \|\hat{X} - X_0\|_F^2 &\leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \left(\langle A - P, \hat{X} - X_0 \rangle + \langle P - Q, \hat{X} - X_0 \rangle \right) \\ &\leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \left(5\sqrt{r} \|A - P\|_2 \|X_0 - \hat{X}\|_F + r(p_{\max} - p_{\min}) \right) \end{aligned} \quad (12)$$

In the dense regime, $\|A - P\|_2 \leq c(\sqrt{\max_{k,\ell} B_{k\ell} n})$ with probability at least $1 - n^{-c_1}$ [24]. Using Lemma 8, with $a = \frac{10\sqrt{r}\|A-P\|_2}{m_{\min} \min_k (p_k - q_k)}$, and $b = \sqrt{\frac{2r(p_{\max} - p_{\min})}{m_{\min} \min_k (p_k - q_k)}}$, with probability at least $1 - n^{-c_1}$, we have:

$$\begin{aligned} \|\hat{X} - X_0\|_F &\leq \frac{10\sqrt{r}\|A - P\|_2}{m_{\min} \min_k (p_k - q_k)} + \sqrt{\frac{2r(p_{\max} - p_{\min})}{m_{\min} \min_k (p_k - q_k)}} \\ &\leq \frac{\sqrt{rnp_{\max}}}{m_{\min} \min_k (p_k - q_k)} \left(c + \sqrt{\frac{2}{r}} \right) \leq \frac{C_G \sqrt{rnp_{\max}}}{m_{\min} \min_k (p_k - q_k)}, \end{aligned}$$

for some positive constant C_G . Since $\|X_0\|_F = \sqrt{r}$, we have the result. \square

5.2 Analysis for low noise covariates

The proof of Theorem 3 is also built on Lemma 3. For kernel concentration, we use the following theorem from Yan and Sarkar [36] characterizing the concentration of the kernel matrix around a suitably specified limit.

Theorem 7 ([36]). *Let $d_{k\ell} = \|\mu_k - \mu_\ell\|$. For $i \in C_k, j \in C_\ell$, define*

$$\tilde{K}(i, j) = \begin{cases} f(d_{k\ell}^2 + \sigma_k^2 + \sigma_\ell^2) & \text{if } i \neq j, \\ f(0) & \text{if } i = j. \end{cases}$$

Then we have $P\left(\|K - \tilde{K}\|_\infty \geq \sqrt{\frac{3 \log n}{\rho d}}\right) \leq \frac{1}{n}$, for some constant $\rho > 0$.

Proof of Theorem 3. Let D' be a diagonal matrix where $D'_{ii} = 1 - f(2\sigma_k^2)$ if $i \in C_k$. $\tilde{K} - D'$ is blockwise constant. When $\min_k \nu_k \geq 0$, this matrix satisfies the condition in Lemma 3, we have

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2\langle K - \tilde{K}, \hat{X} - X_0 \rangle + 2\langle D', \hat{X} - X_0 \rangle}{m_{\min} \min_k \nu_k}$$

By Theorem 7, with probability at least $1 - 1/n$,

$$\langle K - \tilde{K}, \hat{X} - X_0 \rangle \leq n \|\hat{X} - X_0\|_F \cdot \|K - \tilde{K}\|_\infty \leq n \sqrt{\frac{3 \log n}{\rho d}} \|\hat{X} - X_0\|_F \quad (13)$$

The second inequality uses that fact that for $M \in R^{n \times n}$, $\|M\|_F \leq n \|M\|_\infty$. We also have:

$$\langle D', \hat{X} - X_0 \rangle = \sum_k \sum_{i \in C_k} (1 - f(2\sigma_k^2)) (\hat{X}_{ii} - \frac{1}{m_k}) \leq r(f(2\sigma_{\min}^2) - f(2\sigma_{\max}^2)) \leq r \quad (14)$$

Using Lemma 8, with $a = \frac{n \sqrt{12 \log n / \rho d}}{m_{\min} \min_k \nu_k}$, and $b = \sqrt{\frac{2r}{m_{\min} \min_k \nu_k}}$, with probability at least $1 - 1/n$, we have:

$$\begin{aligned} \|\hat{X} - X_0\|_F &\leq a + b = \frac{n \sqrt{12 \log n / \rho d}}{m_{\min} \min_k \nu_k} + \sqrt{\frac{2r}{m_{\min} \min_k \nu_k}} \\ &\leq \frac{n}{m_{\min} \min_k \nu_k} \left(\kappa_1 \sqrt{\frac{\log n}{d}} + \frac{\sqrt{2r m_{\min} \min_k \nu_k}}{n} \right) \\ (\text{Since } r m_{\min} \leq n), &\leq \frac{C_K r \alpha}{\min_k \nu_k} \sqrt{\frac{\log n}{\min(d, n)}}, \end{aligned}$$

for some constants C_K, κ_1 . Since $\|X_0\|_F = \sqrt{r}$, and $n \leq r m_{\max}$, we have the final result. \square

We present a proof sketch of Theorem 1; the details are in the appendix.

Proof sketch of Theorem 1. For dense graph plus high dimensional kernel, we will use $Q = ZBZ^T + \lambda(\tilde{K} - D')$ as the blockwise constant reference matrix, where D' is a diagonal matrix with $D_{ii} = 1 - f(2\sigma_k^2)$ for $i \in C_k$. Let

$$\gamma' = \min_k (\beta(p_k - q_k) + (1 - \beta)\nu_k),$$

where $\beta = \frac{1}{1+\lambda}$. When $\gamma' \geq 0$, by Lemma 3 and Equations 9 and 14,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma} \left(\langle A - P, \hat{X} - X_0 \rangle + \lambda \langle K - \tilde{K}, \hat{X} - X_0 \rangle + r(p_{\max} - p_{\min}) + \lambda r \right)$$

Using Equations 12 and 13 and a little algebra (deferred to the appendix), with probability at least $1 - n^{-c}$, we have, for constants C_G and C_K ,

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{\sqrt{2\alpha^2 r}}{\gamma'} \left(\beta_0 C_G \sqrt{\frac{rp_{\max}}{n}} + (1 - \beta_0) \left(C_K \sqrt{\frac{\log n}{\min(d, n)}} \right) \right)$$

where $\beta_0 = \sqrt{\beta}/(\sqrt{\beta} + \sqrt{1 - \beta})$. □

6 Analysis of sparse graph and high noise co-variates

6.1 Analysis of sparse graph

We start with the following lemma.

Lemma 4 ([17]). *Let $\mathcal{M}_G^+ = \{Z : Z \succeq 0, \text{diag}(Z) \preceq I_n\}$, $A = (a_{ij}) \in \mathbb{R}^{n \times n}$ be a symmetric matrix whose diagonal entries equal 0, and entries above the diagonal are independent random variables satisfying $0 \leq a_{ij} \leq 1$. Let $P = E[A|Z]$. Assume that $\bar{p} := \frac{2}{n(n-1)} \sum_{i < j} \text{Var}(a_{ij}) \geq \frac{9}{n}$. Then, with probability at least $1 - e^{35-n}$, we have*

$$\max_{Z \in \mathcal{M}_G^+} |\langle A - P, Z \rangle| \leq K_G \|A - P\|_{\ell_\infty \rightarrow \ell_1} \leq 3K_G \bar{p}^{1/2} n^{3/2}$$

where K_G is the Grothendieck's constant.

The best known bound for Grothendieck's constant K_G is from [7], where $K_G < \frac{\pi}{2 \log(1+\sqrt{2})} \leq 1.783$.

Proof of Theorem 4. When $\min_k (p_k - q_k) \geq 0$, by Lemma 3 and Equation 9,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min} \min_k (p_k - q_k)} \left(\langle A - P, \hat{X} - X_0 \rangle + r(p_{\max} - p_{\min}) \right)$$

In sparse regime, both $m_{\min}X_0$ and $m_{\min}\hat{X}$ belong to the set \mathcal{M}_G^+ . Let $g = n\bar{p} \geq 9$, applying Lemma 4 we get with probability at least $1 - e^3 5^{-n}$,

$$\begin{aligned}\|\hat{X} - X_0\|_F^2 &\leq \frac{4K_G\|A - P\|_{\ell_\infty \rightarrow \ell_1}}{m_{\min}^2 \min_k(p_k - q_k)} + \frac{2r(p_{\max} - p_{\min})}{m_{\min} \min_k(p_k - q_k)} \\ &\leq \frac{22\sqrt{n^3\bar{p}}}{m_{\min}^2 \min_k(p_k - q_k)} + \frac{2r(p_{\max} - p_{\min})}{m_{\min} \min_k(p_k - q_k)}\end{aligned}$$

Substituting $p_k = a_k/n, q_k = b_k/n$, and using the fact that

$$\frac{2r(p_{\max} - p_{\min})}{m_{\min} \min_k(p_k - q_k)} = \frac{2rm_{\min}(p_{\max} - p_{\min})}{m_{\min}^2 \min_k(p_k - q_k)} \leq \frac{2\max_k a_k}{m_{\min}^2 \min_k(p_k - q_k)} = o(\sqrt{n^3\bar{p}}),$$

Recall that $\alpha := m_{\max}/m_{\min}$, we get with probability tending to 1,

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq \frac{23n^2\sqrt{g}}{rm_{\min}^2 \min_k(a_k - b_k)} \leq \frac{23\alpha^2 r\sqrt{g}}{\min_k(a_k - b_k)}.$$

□

6.2 Analysis of high noise kernel

The details of the following proof sketch is deferred to the supplementary material.

Proof sketch of Theorem 5. Recall that by definition, for $i \in C_k$, $Y_i - \mu_k$ is sub-gaussian random vector with sub-gaussian norm $\frac{\psi_k}{\sqrt{d}}$. Using the following concentration inequality from [20] for sub-gaussian random vectors, we have:

$$\text{For } i \in C_k, P(\|Y_i - \mu_k\|_2^2 > \psi_k^2(1 + 2\sqrt{t/d} + 2t/d)) \leq e^{-t}$$

We take $t = c_k^2 d$ for $c_k \geq 1$. Since $1 + 2c_k + 2c_k^2 \leq 5c_k^2$ for $c_k \geq 1$, we get $P(\|X - \mathbb{E}X\|^2 \leq 5c_k^2\psi_k^2) \geq 1 - \exp(-c_k^2 d)$. Let $\Delta_k = \sqrt{5c_k}\psi_k$, we can divide the nodes into “good nodes” (those close to their population mean) \mathcal{S}_k and the rest as follows:

$$\mathcal{S}_k = \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}, \quad \mathcal{S} = \cup_{k=1}^r \mathcal{S}_k \quad (15)$$

Let $m_c^{(k)} = m_k - |\mathcal{S}_k|$. We want to bound $m_c^{(k)}$ with high probability. Note that $m_c^{(k)} = \sum_{i \in C_k} \mathbf{1}(\|Y_i - \mu_k\| \geq \Delta_k)$ is a sum of i.i.d random variables. Therefore, using the Hoeffding bound for each cluster and union bound over all clusters we get:

$$P\left(m_c \geq \sum_k m_k e^{-c_k^2 d} + \sum_k \sqrt{m_k \log m_k / 2}\right) \leq \sum_k \frac{1}{m_k} \quad (16)$$

Now define

$$(K_I)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i, j \in C_k \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \neq \ell \end{cases} \quad (17)$$

By construction, all diagonal blocks of K_I are blockwise constant and the off-diagonal blocks are upper bounded by $f(d_{k\ell} - \Delta_k - \Delta_\ell)$. Let $\nu_k = f(2\Delta_k) -$

$\max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$, and $\gamma = \min_k \nu_k$. When $\gamma \geq 0$, by Lemma 3 and Grothendieck's inequality, we have

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma} \langle K - K_I, \hat{X} - X_0 \rangle \leq \frac{4K_G}{m_{\min}^2\gamma} \|K - K_I\|_{\ell_\infty \rightarrow \ell_1} \quad (18)$$

Now it remains to bound the $\ell_\infty \rightarrow \ell_1$ norm of $K - K_I$. Note that if $i \in S_k, j \in S_\ell, k \neq \ell$, then by a simple use of triangle inequality we have $K_{ij} \leq f(d_{k\ell} - \Delta_k - \Delta_\ell)$, so $K_{ij} = (K_I)_{ij}$; and if $i, j \in S_k$, then $K_{ij} \geq f(2\Delta_k)$.

$$\begin{aligned} \|K - K_I\|_{\ell_\infty \rightarrow \ell_1} &= \max_{x, y \in \{\pm\}^n} \sum_{i, j} x_i y_j (K_{ij} - (K_I)_{ij}) \\ &\leq \max_{x, y \in \{\pm\}^n} \sum_{i, j \in S} x_i y_j (K_{ij} - (K_I)_{ij}) + \max_{x, y \in \{\pm\}^n} \sum_{i \notin S \cup j \notin S} x_i y_j (K_{ij} - (K_I)_{ij}) \\ &\stackrel{(i)}{\leq} \max_{x, y \in \{\pm\}^n} \sum_{i, j \in S} x_i y_j (K_{ij} - (K_I)_{ij}) + 2m_c n \\ &\stackrel{(ii)}{=} \max_{x, y \in \{\pm\}^n} \sum_k \sum_{i, j \in S_k} x_i y_j (K_{ij} - f(2\Delta_k)) + 2m_c n \leq \sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n \end{aligned} \quad (19)$$

where (i) is due to $|K_{ij} - (K_I)_{ij}| \leq 1$, and (ii) comes from the definition of K_I . Now Equation 31 leads to:

$$\|\hat{X} - X_0\|_F^2 \leq \frac{4K_G (\sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n)}{m_{\min}^2 \gamma} \quad (20)$$

Recall that $f(x) = \exp(-\eta x^2)$. For simplicity, we assume $c_k = c_0$. For some $\phi > 0$ to be chosen later, we define:

$$c_0 = \sqrt{\log \left(\frac{d_{\min}^2}{\psi_{\max}^2} \right)} / d, \quad \xi = \frac{d_{\min}}{2\sqrt{5}c_0\psi_{\max}} - 1, \quad \eta = \frac{\phi}{20c_0^2\psi_{\max}^2}. \quad (21)$$

Thus, when $d_{\min} := \min_{k\ell} d_{k\ell} > 4\sqrt{5}c_0\psi_{\max}$,

$$\begin{aligned} \gamma &\geq f(2\sqrt{5}c_0\psi_{\max}) - f(d_{\min} - 2\sqrt{5}c_0\psi_{\max}) = \exp(-\phi) - \exp(-\phi\xi^2). \\ 1 - f(2\Delta_k) &\leq 1 - f(2\sqrt{5}c_0\psi_{\max}) = 1 - \exp(\phi) \end{aligned} \quad (22)$$

Denoting $\alpha = \frac{m_{\max}}{m_{\min}}$, Equations 29, 33 and 37 gives:

$$\|\hat{X} - X_0\|_F^2 \leq 4K_G r \alpha^2 \left(\underbrace{\frac{(1 - \exp(-\phi) + 2r\psi_{\max}^2/d_{\min}^2)}{\exp(-\phi) - \exp(-\phi\xi^2)}}_A + \underbrace{\frac{r\sqrt{\log m_{\max}/2m_{\min}}}{\exp(-\phi) - \exp(-\phi\xi^2)}}_B \right) \quad (23)$$

We will first bound part (A).

$$(A) = \frac{\exp(\phi) - 1 + \exp(\phi) \frac{2r\psi_{\max}^2}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \stackrel{(i)}{\leq} \frac{\phi + \frac{\phi^2}{2} \exp(\phi) + \exp(\phi) \frac{2r\psi_{\max}^2}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \quad (24)$$

where (i) uses the Mean value theorem: for $e^x - 1 \leq x + e^y x^2/2$ for $y \in [0, x]$. If $\frac{d_{\min}}{\psi_{\max}} > \max\{1, \frac{180}{d}\}$, using the fact that $\log x \leq \sqrt{x}$, we have:

$$\frac{d_{\min}^2}{\psi_{\max}^2} > \frac{180}{d} \frac{d_{\min}}{\psi_{\max}} > \frac{180}{d} \log\left(\frac{d_{\min}^2}{\psi_{\max}^2}\right) = 180c_0^2.$$

Therefore, $d_{\min} > 4\sqrt{5}c_0\psi_{\max}$ and hence $\gamma > 0$. Also, Equation 34 shows that with this condition on d_{\min} , $\xi \geq 2$. In order to make the separation large, we need to pick ϕ such that for large ξ , ϕ is small, but $\phi\xi^2$ is large. So we pick $\phi = \frac{\log \xi}{\xi^2}$. Since $\frac{\log x}{x^2}$ is monotonically decreasing when $x > 2$, $\phi < \log 2/2^2$. Using this and some algebra, we can show that $1 - \exp(\phi - \phi\xi^2) \geq 1 - \exp(\phi\xi^2/4 - \phi\xi^2) \geq .4$. Now Equation 36 yields:

$$(A) \leq \frac{\phi + 1.2\left(\frac{\phi^2}{2} + \frac{2r\psi_{\max}^2}{d_{\min}^2}\right)}{.4} \stackrel{(ii)}{\leq} c \frac{\psi_{\max}^2}{d_{\min}^2} \log\left(\frac{d_{\min}}{\psi_{\max}}\right) + \frac{3r\psi_{\max}^2}{d_{\min}^2}, \quad (25)$$

for some constant c . To get (ii), note that

$$\frac{\log \xi}{\xi^2} \leq \frac{\log(\xi + 1)}{\xi^2} \leq \frac{2.25 \log(\xi + 1)}{(\xi + 1)^2}, \forall \xi > 2$$

Using the fact that the function $\frac{\log x}{x^2}$ is monotonically decreasing when $x > 2$, we see that $\phi < \log 2/2^2$ and

$$\gamma \geq \exp(-\phi)(1 - \exp(\phi(1 - \xi^2))) \geq .3 \quad (26)$$

Finally, we bound (B) in Equation 35 using Equation 37.

$$(B) = \frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)} \leq c_1 r \sqrt{\frac{\log m_{\max}}{m_{\max}}} \quad (27)$$

for some constant $c_1 > 0$. Substituting Equations 25 and 27 into Equation 35, we have

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 \max\left(\frac{\psi_{\max}^2}{d_{\min}^2} \max\left\{\log\left(\frac{d_{\min}}{\psi_{\max}}\right), r\right\}, r\sqrt{\frac{\log m_{\max}}{m_{\max}}}\right)$$

□

6.3 Analysis for sparse graph plus high noise kernel

Proof of Theorem 6. Let K_I be defined as in Equation 30. Let $\gamma = \min_k(p_k - q_k + \lambda(f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)))$. When $\gamma \geq 0$, lemma 3 with $Q = ZBZ^T + \lambda K_I$, we have

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min}\gamma} \left(\langle A - P, \hat{X} - X_0 \rangle + r(p_{\max} - p_{\min}) + \lambda \langle K - K_I, \hat{X} - X_0 \rangle \right)$$

Now by Grothendieck's inequality on both $\langle A - P, \hat{X} - X_0 \rangle$ and $\langle K - K_I, \hat{X} - X_0 \rangle$, one gets,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2K_G}{m_{\min}^2\gamma} (2\|A - P\|_{\ell_\infty \rightarrow \ell_1} + r(p_{\max} - p_{\min}) + 2\lambda\|K - K_I\|_{\ell_\infty \rightarrow \ell_1})$$

By Lemma 4 and Equation 32,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{4K_G}{m_{\min}^2 \gamma} \left(6\sqrt{n^3 \bar{p}} + \lambda \left(2m_c n + \sum_k m_k^2 (1 - f(2\Delta_k)) \right) \right)$$

Recall that for the sparse graph, $p_k = a_k/n, q_k = b_k/n, g = \bar{p}/n$. Using $\lambda = \ell/n, m_k = n\pi_k, m_{\min} = n\pi_{\min}$, and $\pi_0 := \sum_k (m_k \exp(-c_k^2 d) + \sqrt{m_k \log m_k / 2})/n$ in conjunction with Equation 29, we get:

$$\|\hat{X} - X_0\|_F^2 \leq 4K_G \frac{6\sqrt{g} + \ell (2\pi_0 + \sum_k \pi_k^2 (1 - f(2\Delta_k)))}{\pi_{\min}^2 \min_k (a_k - b_k + \ell \nu_k)}$$

□

7 Discussion

In this paper, we propose a semidefinite relaxation for clustering a network with covariates. We theoretically establish that if there is a cluster that cannot be distinguished using one source, but the other source can distinguish it, then the proposed SDP is able to separate it correctly with high probability. Furthermore, our proof technique is simple and it unifies the analysis for sub-gaussian mixtures in high or low dimensional spaces, strongly or weakly assortative, sparse or dense networks with equal or unequal sized clusters. We demonstrate the performance of our methodology on simulated and real networks, and show that it in general performs better than other state-of-the-art methods.

Acknowledgements

We thank Arash Amini and Yuan Zhang for generously sharing their code. We are also grateful to Soumendu Mukherjee and Peter J. Bickel for their valuable comments on our manuscript.

Appendix

A Sub-gaussian random vectors

In our analysis, we make use of some useful properties of sub-gaussian random vectors. In this section, we collect some of their properties. First of all, a sub-gaussian random variable is defined by the following equivalent properties. More discussions on this topic can be found in [32].

Lemma 5 ([32]). *The sub-gaussian norm of X is denoted by $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}$. A random vector $X \in \mathbb{R}^n$ is defined to be sub-gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$ with sub-gaussian norm $\|X\|_{\psi_2} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_2}$.*

Every sub-gaussian random variable X satisfies:

$$(1) \ P(|X| > t) \leq \exp(1 - ct^2 / \|X\|_{\psi_2}^2) \text{ for all } t \geq 0;$$

- (2) (Rotation invariance) Consider a finite number of independent centered sub-gaussian random variables X_i . Then $\sum_i X_i$ is also a centered sub-gaussian random variable. Moreover, $\|\sum_i X_i\|_{\psi_2}^2 \leq C \sum_i \|X_i\|_{\psi_2}^2$.
- (3) Let X_1, \dots, X_n be independent centered sub-gaussian random variables. Then $X = (X_1, X_2, \dots, X_n)$ is a centered sub-gaussian random vector in \mathbb{R}^n and $\|X\|_{\psi_2} \leq C \max_i \|X_i\|_{\psi_2}$.

A random variable is sub-exponential if the following equivalent properties hold with parameters $K_i > 0$ differing from each other by at most an absolute constant factor: (1) $P(|X| > t) \leq \exp(1 - t/K_1)$ for all $t \geq 0$; (2) $(\mathbb{E}|X|)^{1/p} \leq K_2 p$ for all $p \geq 1$; (3) $\mathbb{E} \exp(X/K_3) \leq e$. The square of sub-gaussian random variable is sub-exponential.

Lemma 6 ([32]). A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover, $\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$.

B Detailed analysis for dense graph and low noise covariates

B.1 Accompanying lemmas for analyzing dense graph

Lemma 7. For any X that satisfies $X \succeq 0, X \geq 0, X\mathbf{1} = \mathbf{1}$, we have $\|X\|_F^2 \leq \text{trace}(X)$.

Proof of Lemma 7. We first show that for all such X , the eigenvalues of X are in $[0, 1]$. Let v_i be the eigenvector of X corresponding to the i^{th} largest eigenvalue λ_i . Since X is positive semi-definite, $\lambda_i \geq 0, \forall i$. Without loss of generality, let $i^* = \arg \max_i |v_1(i)|$, i.e. be the index of the entry with the largest absolute value of v_1 . Since $Xv_1 = \lambda_1 v_1$, and $\sum_j X_{ij} = 1, X_{ij} \geq 0$, we have:

$$|\lambda_1 v_1(i^*)| = |\sum_j X_{i^*j} v_1(j)| \leq \sum_j X_{i^*j} |v_1(j)| \leq |v_1(i^*)|.$$

Therefore $|\lambda_1| \leq 1$.

$$\|X\|_F^2 = \sum_i \lambda_i^2 \leq \sum_i \lambda_i = \text{trace}(X)$$

□

Lemma 8. If $x^2 - ax - b^2 \leq 0$, for some $a, b > 0$, then $x \leq a + b$.

Proof. Solving quadratic inequality we have $x \leq \frac{a + \sqrt{a^2 + 4b^2}}{2}$. Using $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$, $x \leq \frac{a + a + 2\sqrt{b^2}}{2} = a + b$. □

B.2 Detailed analysis for dense graph plus low noise covariates

Lemma 9. If $x^2 - (\beta a_1 + (1 - \beta)a_2)x - (\beta b_1 + (1 - \beta)b_2) \leq 0$, where $\beta \in [0, 1]$, $x \leq (\sqrt{\beta}(a_1 + \sqrt{b_1}) + \sqrt{1 - \beta}(a_2 + \sqrt{b_2}))$.

Proof. Using Lemma 8 we see that:

$$\begin{aligned} x &\leq \beta a_1 + (1 - \beta)a_2 + \sqrt{\beta b_1 + (1 - \beta)b_2} \\ &\leq \sqrt{\beta a_1^2 + (1 - \beta)a_2^2} + \sqrt{\beta b_1 + (1 - \beta)b_2} \\ &\leq \sqrt{\beta}(a_1 + \sqrt{b_1}) + \sqrt{1 - \beta}(a_2 + \sqrt{b_2}) \end{aligned}$$

□

Theorem 8 (Dense graph plus low noise kernel). *Let $\gamma' = \min_k \left(\frac{p_k - q_k}{1 + \lambda} + \frac{\lambda}{1 + \lambda} \nu_k \right)$, where ν_k is defined as in Equation 5. If $\nu_k \geq 0$, we have:*

$$\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} \leq \frac{\sqrt{2\alpha^2 r}}{\gamma'} \left(\frac{1}{1 + \sqrt{\lambda}} C_G \sqrt{\frac{r p_{\max}}{n}} + \frac{\sqrt{\lambda}}{1 + \sqrt{\lambda}} C_K \sqrt{\frac{\log n}{\min(d, n)}} \right)$$

Proof of Theorem 8. For dense graph plus low noise kernel, we will use $Q + \lambda(\tilde{K} - D')$ as the blockwise constant reference matrix, where D' is a diagonal matrix with $D_{ii} = 1 - f(2\sigma_k^2)$ for $i \in C_k$. Let

$$\gamma' = \min_k (\beta(p_k - q_k) + (1 - \beta)\nu_k),$$

where $\beta = \frac{1}{1 + \lambda}$. By Lemma 3 and Equations 9, 14, when $\gamma' \geq 0$,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min} \gamma'} \left(\langle A - P, \hat{X} - X_0 \rangle + \lambda \langle K - \tilde{K}, \hat{X} - X_0 \rangle + r(p_{\max} - p_{\min}) + \lambda r \right)$$

Use Equations 12 and 13, with probability at least $1 - n^{-c}$,

$$\begin{aligned} &\|\hat{X} - X_0\|_F^2 \\ &\leq \frac{2 \left(\beta \langle A - P, \hat{X} - X_0 \rangle + (1 - \beta) \langle K - \tilde{K}, \hat{X} - X_0 \rangle + \beta r(p_{\max} - p_{\min}) + (1 - \beta)r \right)}{m_{\min} \gamma'} \\ &\leq \frac{2}{m_{\min} \gamma'} \left(\left(5\beta\sqrt{r}\|A - P\|_2 + (1 - \beta)n\sqrt{\frac{3\log n}{\rho d}} \right) \|\hat{X} - X_0\|_F + \beta r(p_{\max} - p_{\min}) + (1 - \beta)r \right) \end{aligned}$$

Using Lemma 9 we see that:

$$\begin{aligned} &\|\hat{X} - X_0\|_F \\ &\leq \sqrt{\beta} \left(\frac{10\sqrt{r}\|A - P\|_2}{m_{\min} \gamma'} + \sqrt{\frac{2r(p_{\max} - p_{\min})}{m_{\min} \gamma'}} \right) \\ &\quad + \sqrt{1 - \beta} \left(\frac{2n}{m_{\min} \gamma'} \sqrt{\frac{3\log n}{\rho d}} + \sqrt{\frac{2r}{m_{\min} \gamma'}} \right) \\ &\leq \sqrt{\beta} \frac{C_G \sqrt{r n p_{\max}}}{m_{\min} \gamma'} + \sqrt{1 - \beta} \frac{n}{m_{\min} \gamma'} \left(\kappa_1 \sqrt{\frac{\log n}{d}} + \kappa_2 \sqrt{\frac{1}{n}} \right) \\ &\leq \frac{n}{m_{\min} \gamma'} \left(\sqrt{\beta} C_G \sqrt{\frac{r p_{\max}}{n}} + \sqrt{1 - \beta} \left(\kappa_1 \sqrt{\frac{\log n}{d}} + \kappa_2 \sqrt{\frac{1}{n}} \right) \right) \end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\|\hat{X} - X_0\|_F}{\|X_0\|_F} &\leq \frac{\sqrt{\alpha^2 r}}{\gamma'} \left(\sqrt{\beta} C_G \sqrt{\frac{r p_{\max}}{n}} + \sqrt{1 - \beta} \left(\kappa_1 \sqrt{\frac{\log n}{d}} + \kappa_2 \sqrt{\frac{1}{n}} \right) \right) \\ &\leq \frac{\sqrt{2\alpha^2 r}}{\gamma'} \left(\beta_0 C_G \sqrt{\frac{r p_{\max}}{n}} + (1 - \beta_0) \left(C_K \sqrt{\frac{\log n}{\min(d, n)}} \right) \right)\end{aligned}$$

where $\beta_0 = \sqrt{\beta}/(\sqrt{\beta} + \sqrt{1 - \beta})$.

□

C Detailed analysis of high noise covariates

Theorem 9 (High noise kernel). *Let \hat{X} be the optimal solution of **SDP-cov**, $\alpha = \frac{m_{\max}}{m_{\min}}$. If $\frac{d_{\min}}{\psi_{\max}} > \max\{1, \frac{180}{d}\}$, then with properly chosen η , with probability at least $1 - \sum_k \frac{1}{m_k}$,*

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq C \alpha^2 \frac{\psi_{\max}^2}{d_{\min}^2} \max \left\{ \log \left(\frac{d_{\min}}{\psi_{\max}} \right), r \right\}$$

Proof of Theorem 9. Recall that by definition, for $i \in C_k$, $Y_i - \mu_k$ is sub-gaussian random vector with sub-gaussian norm $\frac{\psi_k}{\sqrt{d}}$. Using the following concentration inequality from [20] for sub-gaussian random vectors, we have:

$$\text{For } i \in C_k, P(\|Y_i - \mu_k\|_2^2 > \psi_k^2(1 + 2\sqrt{t/d} + 2t/d)) \leq e^{-t}$$

We take $t = c_k^2 d$ for $c_k \geq 1$. Since $1 + 2c_k + 2c_k^2 \leq 5c_k^2$ for $c_k \geq 1$, we get $P(\|X - \mathbb{E}X\|^2 \leq 5c_k^2 \psi_k^2) \geq 1 - \exp(-c_k^2 d)$. Let $\Delta_k = \sqrt{5}c_k \psi_k$, we can divide the nodes into “good nodes” (those close to their population mean) \mathcal{S}_k and the rest as follows:

$$\mathcal{S}_k = \{i \in C_k : \|Y_i - \mu_k\| \leq \Delta_k\}, \quad \mathcal{S} = \cup_{k=1}^r \mathcal{S}_k \quad (28)$$

Let $m_c^{(k)} = m_k - |\mathcal{S}_k|$. We want to bound $m_c^{(k)}$ with high probability. Note that $m_c^{(k)} = \sum_{i \in C_k} \mathbf{1}(\|Y_i - \mu_k\| \geq \Delta_k)$ is a sum of i.i.d random variables. Therefore, using the Hoeffding bound we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq m_k \delta\right) \leq \exp(-2m_k \delta^2)$$

Using $\delta = \sqrt{\log m_k / 2m_k}$, we have:

$$P\left(m_c^{(k)} - m_k P(i \notin \mathcal{S}_k) \geq \sqrt{m_k \log m_k / 2}\right) \leq \frac{1}{m_k}$$

Since $P(i \notin \mathcal{S}_k) \leq \exp(-c_k^2 d)$, we have:

$$P\left(m_c^{(k)} \geq m_k \exp(-c_k^2 d) + \sqrt{m_k \log m_k / 2}\right) \leq \frac{1}{m_k}$$

Finally, using union bound over all clusters we get:

$$P\left(m_c \geq \sum_k m_k e^{-c_k^2 d} + \sum_k \sqrt{m_k \log m_k / 2}\right) \leq \sum_k \frac{1}{m_k} \quad (29)$$

Now define

$$(K_I)_{ij} = \begin{cases} f(2\Delta_k), & \text{if } i, j \in C_k \\ \min\{f(d_{k\ell} - \Delta_k - \Delta_\ell), K_{ij}\}, & \text{if } i \in C_k, j \in C_\ell, k \neq \ell \end{cases} \quad (30)$$

By Lemma 3, all diagonal blocks are blockwise constant and the off-diagonal blocks are upper bounded by $f(d_{k\ell} - \Delta_k - \Delta_\ell)$. Let $\nu_k = f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)$, and $\gamma = \min_k \nu_k$. If $\nu_k \geq 0$, we have

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2}{m_{\min} \gamma} \langle K - K_I, \hat{X} - X_0 \rangle$$

Apply Grothendieck's inequality,

$$\|\hat{X} - X_0\|_F^2 \leq \frac{2K_G}{m_{\min}^2 \gamma} \|K - K_I\|_{\ell_\infty \rightarrow \ell_1} \quad (31)$$

Now it remains to bound the $\ell_\infty \rightarrow \ell_1$ norm of $K - K_I$. Note that if $i \in S_k, j \in S_\ell, k \neq \ell$, then by a simple use of triangle inequality we have $K_{ij} \leq f(d_{k\ell} - \Delta_k - \Delta_\ell)$, so $K_{ij} = (K_I)_{ij}$; and if $i, j \in S_k$, then $K_{ij} \geq f(2\Delta_k)$.

$$\begin{aligned} \|K - K_I\|_{\ell_\infty \rightarrow \ell_1} &= \max_{x, y \in \{\pm\}^n} \sum_{i, j} x_i y_j (K_{ij} - (K_I)_{ij}) \\ &\leq \max_{x, y \in \{\pm\}^n} \sum_{i, j \in \mathcal{S}} x_i y_j (K_{ij} - (K_I)_{ij}) + \max_{x, y \in \{\pm\}^n} \sum_{i \notin \mathcal{S} \cup j \notin \mathcal{S}} x_i y_j (K_{ij} - (K_I)_{ij}) \\ &\stackrel{(i)}{\leq} \max_{x, y \in \{\pm\}^n} \sum_{i, j \in \mathcal{S}} x_i y_j (K_{ij} - (K_I)_{ij}) + 2m_c n \\ &\stackrel{(ii)}{=} \max_{x, y \in \{\pm\}^n} \sum_k \sum_{i, j \in S_k} x_i y_j (K_{ij} - f(2\Delta_k)) + 2m_c n \\ &\leq \sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n \end{aligned} \quad (32)$$

where (i) is due to $|K_{ij} - (K_I)_{ij}| \leq 1$, and (ii) comes from the definition of K_I . Now Equation 31 follows as

$$\begin{aligned} \|\hat{X} - X_0\|_F^2 &\leq \frac{4K_G (\sum_k m_k^2 (1 - f(2\Delta_k)) + 2m_c n)}{m_{\min}^2 \gamma} \\ &= \frac{4K_G}{m_{\min}^2} \sum_k \left(m_k^2 \frac{1 - f(2\Delta_k)}{\gamma} + 2m_k n e^{-c_k^2 d} / \gamma \right) + \frac{\sqrt{2}K_G n}{m_{\min}^2 \gamma} \sum_k \sqrt{m_k \log m_k} \end{aligned} \quad (33)$$

Recall that $f(x) = \exp(-\eta x^2)$, and $\gamma = \min_k \{f(2\Delta_k) - \max_{\ell \neq k} f(d_{k\ell} - \Delta_k - \Delta_\ell)\}$. For simplicity, we assume $c_k = c_0$. We take $c_0 = \sqrt{\log\left(\frac{d_{\min}^2}{\psi_{\max}^2}\right)/d}$ and the scale

parameter $\eta = \frac{\phi}{20c_0^2\psi_{\max}^2}$, for some $\phi > 0$, which will be chosen later. Furthermore, we also define

$$\xi = \frac{d_{\min}}{2\sqrt{5}c_0\psi_{\max}} - 1. \quad (34)$$

If $\xi > 1$, then $d_{\min} > 4\sqrt{5}c_0\psi_{\max}$, and hence $\gamma > 0$. Also, since $\eta(d_{\min} - 2\sqrt{5}c_0\psi_{\max})^2 = \phi\xi^2$, $\forall k, \ell \in [r]$, if $d_{\min} := \min_{k\ell} d_{k\ell} > 4\sqrt{5}c_0\psi_{\max}$, then

$$\gamma \geq f(2\sqrt{5}c_0\psi_{\max}) - f(d_{\min} - 2\sqrt{5}c_0\psi_{\max}) = \exp(-\phi) - \exp(-\phi\xi^2).$$

and

$$1 - f(2\Delta_k) \leq 1 - f(2\sqrt{5}c_0\psi_{\max}) = 1 - \exp(\phi)$$

Denoting $\alpha = \frac{m_{\max}}{m_{\min}}$,

$$\begin{aligned} \|\hat{X} - X_0\|_F^2 &\leq 4K_G r \alpha^2 \cdot \frac{1 - f(2\sqrt{5}c_0\psi_{\max}) + 2r \exp(-c_0^2 d)}{\gamma} + \frac{2\sqrt{2}K_G m_{\max} r^2 \sqrt{m_{\max} \log m_{\max}}}{\gamma m_{\min}^2} \\ &\leq \frac{4K_G r \alpha^2}{\gamma} \left(1 - \exp(-\phi) + \frac{2r\psi_{\max}^2}{d_{\min}^2} + r\sqrt{\log m_{\max}/2m_{\max}} \right) \\ &\leq 4K_G r \alpha^2 \left(\underbrace{\frac{(1 - \exp(-\phi) + 2r\psi_{\max}^2/d_{\min}^2)}{\exp(-\phi) - \exp(-\phi\xi^2)}}_A + \underbrace{\frac{r\sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)}}_B \right) \end{aligned} \quad (35)$$

We will first bound part (A).

$$(A) = \frac{\exp(\phi) - 1 + \exp(\phi) \frac{2r\psi_{\max}^2}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \stackrel{(i)}{\leq} \frac{\phi + \frac{\phi^2}{2} \exp(\phi) + \exp(\phi) \frac{2r\psi_{\max}^2}{d_{\min}^2}}{1 - \exp(\phi - \phi\xi^2)} \quad (36)$$

where (i) uses the Mean value theorem: for $e^x - 1 \leq x + e^y x^2/2$ for $y \in [0, x]$. If $\frac{d_{\min}}{\psi_{\max}} > \max\{1, \frac{180}{d}\}$, using the fact that $\log x \leq \sqrt{x}$, we have:

$$\frac{d_{\min}^2}{\psi_{\max}^2} > \frac{180}{d} \frac{d_{\min}}{\psi_{\max}} > \frac{180}{d} \log \left(\frac{d_{\min}^2}{\psi_{\max}^2} \right) = 180c_0^2.$$

Using Equation 34, we see that $\xi > \frac{\sqrt{180}}{2\sqrt{5}} - 1 = 2$, and hence $\gamma > 0$. Now we pick $\phi = \frac{\log \xi}{\xi^2}$.

Now we will use this to obtain a lower bound on $1 - \exp(\phi - \phi\xi^2)$. Since $\xi \geq 2$, we have $\xi^2/4 \geq 1$. Hence

$$\begin{aligned} 1 - \exp(\phi - \phi\xi^2) &\geq 1 - \exp(\phi\xi^2/4 - \phi\xi^2) \\ &= 1 - \exp(-\phi 3\xi^2/4) = 1 - \exp(-3 \log \xi/4) = 1 - \xi^{-3/4} \\ &\geq 1 - 2^{-3/4} = .4 \end{aligned}$$

Using the fact that the function $\frac{\log x}{x^2}$ is monotonically decreasing when $x > 2$, we see that $\phi < \log 2/2^2$ and $\exp(\phi) \leq 1.2$. Furthermore,

$$\gamma \geq \exp(-\phi)(1 - \exp(\phi(1 - \xi^2))) \geq .3 \quad (37)$$

Now Equation 36 yields:

$$\begin{aligned} (A) &\leq \frac{\phi + 1.2 \left(\frac{\phi^2}{2} + \frac{2r\psi_{\max}^2}{d_{\min}^2} \right)}{.4} \leq \frac{c \log \xi}{\xi^2} + \frac{3r\psi_{\max}^2}{d_{\min}^2} \\ &\stackrel{(ii)}{\leq} \frac{c' \log(\xi + 1)}{(\xi + 1)^2} + \frac{3r\psi_{\max}^2}{d_{\min}^2} \leq c'' \frac{\psi_{\max}^2}{d_{\min}^2} \log \left(\frac{d_{\min}}{\psi_{\max}} \right) + \frac{3r\psi_{\max}^2}{d_{\min}^2}, \end{aligned}$$

for some constant c . To get (ii), note that

$$\frac{\log \xi}{\xi^2} \leq \frac{\log(\xi + 1)}{\xi^2} \leq \frac{2.25 \log(\xi + 1)}{(\xi + 1)^2}, \forall \xi > 2$$

Finally, we bound (B) in Equation 35 using Equation 37.

$$(B) = \frac{r \sqrt{\log m_{\max}/2m_{\max}}}{\exp(-\phi) - \exp(-\phi\xi^2)} \leq c_1 r \sqrt{\frac{\log m_{\max}}{m_{\max}}}$$

for some constant $c_1 > 0$. Putting pieces together, we have

$$\frac{\|\hat{X} - X_0\|_F^2}{\|X_0\|_F^2} \leq C\alpha^2 \max \left(\frac{\psi_{\max}^2}{d_{\min}^2} \max \left\{ \log \left(\frac{d_{\min}}{\psi_{\max}} \right), r \right\}, r \sqrt{\frac{\log m_{\max}}{m_{\max}}} \right)$$

□

D Implementation

We use augmented Lagrangian method of multipliers to implement **SDP-comb**. Let $\Phi_i = e_i 1_n^T + 1_n e_i^T$, then the constraint $X \mathbf{1}_n = \mathbf{1}_n$ is equivalent to $\langle \Phi_i, X \rangle = 2$. The problem can be reformulated as

$$\begin{aligned} \min_X \quad & -\langle A, X \rangle + 1(L(X) = b) + 1(Y \succeq 0) + 1(0 \leq Z \leq 1/m_{\min}), \\ \text{s.t.} \quad & X = Y, X = Z \end{aligned}$$

where $1(\cdot)$ represents the indicator function, $L(X)_i = \langle \Phi_i, X \rangle, b_i = 2, \forall i \in [n]$, $L(X)_{n+1} = \langle I_n, X \rangle, b_{n+1} = r$.

The pseudo code is summarized in Algorithm 1. In line 5, Π_{S^+} computes the projection into the space of semidefinite matrices. To be more explicit, let $X \in \mathbb{R}^{n \times n}$ be a matrix with eigendecomposition $X = P\Lambda P^T$, $\Lambda = \text{diag}(\lambda_1 \cdots \lambda_n)$. Now we have

$$\Pi_{S^+}(X) = P \text{diag}(\max(0, \lambda_1), \cdots, \max(0, \lambda_n)) P^T$$

The projection on line 3 can be calculated in closed form.

$$\Pi_L(X) = X - L^*(LL^*)^{-1}(L(X) - b)$$

$$\text{where } (LL^*)^{-1} = \begin{bmatrix} \frac{1}{2n} (I_n - \frac{n-2}{n(2n-2)} E) & \frac{1}{n(2-2n)} \mathbf{1}_n \\ \frac{1}{2n(1-n)} \mathbf{1}_n^T & \frac{1}{n-1} \end{bmatrix}.$$

The stopping criterion is $\|X^{k+1} - X^k\|/\|X^k\|$ less than some tolerance or it achieves a certain number of iterations.

Algorithm 1 ADMM for **SDP-comb**

Require: Network A , node covariate matrix Y , tuning parameter λ, ρ .

- 1: Compute kernel matrix K where $K(i, j) = f(\|Y_i - Y_j\|_2^2)$;
 - 2: **while** not converge **do**
 - 3: $X^{(k+1)} = \Pi_L(\frac{1}{2}(Z^k - U^k + Y^k - V^k) + \frac{1}{\rho}(A + \lambda K))$;
 - 4: $Z^{(k+1)} = \max(0, X^{(k+1)} + U^k)$, $Z^{(k+1)} = \min(Z^{(k+1)}, 1/m_{\min})$;
 - 5: $Y^{(k+1)} = \Pi_{S^+}(X^{(k+1)} + V^k)$;
 - 6: $U^{(k+1)} = U^k + X^{(k+1)} - Z^{(k+1)}$;
 - 7: $V^{(k+1)} = V^k + X^{(k+1)} - Y^{(k+1)}$;
 - 8: **end while**
 - 9: Return X^k .
-

E Additional experimental results

In this simulation, we fix the block connection probability matrix B , and the distances between covariate centers, and report NMI on the Y-axis with increasing n on the X-axis. We compare **SDP-net**, **SDP-comb**, **SDP-cov**, ACASC and RSC. Our tuning parameter is chosen by the method described in Section 4.1. For ACASC we use the method outlined by the authors for finding the tuning parameter.

For these simulations, we illustrate the case of orthogonal information from two sources, i.e. the network and the covariates, using the following example. We consider the following weakly assortative SBM whose probability matrix is $B = \begin{bmatrix} 0.2 & 0.16 & 0.08 \\ 0.16 & 0.2 & 0.1 \\ 0.08 & 0.1 & 0.12 \end{bmatrix}$. We generate the centers where $\mu_1 = (0.4, 0, 0, 0 \cdots, 0)$, $\mu_2 = (0, 0.1, 0, 0 \cdots, 0)$, $\mu_3 = (0, 0, 0.1, 0 \cdots, 0)$ and $\sigma = 1$. In this example, the network cannot separate out clusters one and two well, whereas the covariates can. On the other hand, clusters two and three are not well separated in the covariate space, while they are well separated using the network parameters. Figure 6(a) presents the comparison between the true and recovered structure. Both **SDP-comb** and ACASC outperform the other methods by a large margin since they combine the two sources of information. Furthermore, **SDP-comb** outperforms ACASC.

In Figure 6(b), we use another example to demonstrate a situation where both the covariates and the network can lead to consistent clustering. For this, we only compare the NMI of **SDP-net**, **SDP-comb** and **SDP-cov** as n is increased on the X-axis. We show that while all of them achieve perfect clustering eventually, **SDP-comb** outperforms both by a large margin by combining both.

References

- [1] Amini, A. A., Chen, A., Bickel, P. J., Levina, E., et al. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.
- [2] Amini, A. A. and Levina, E. (2014). On semidefinite relaxations for the block model. *arXiv preprint arXiv:1406.5647*.

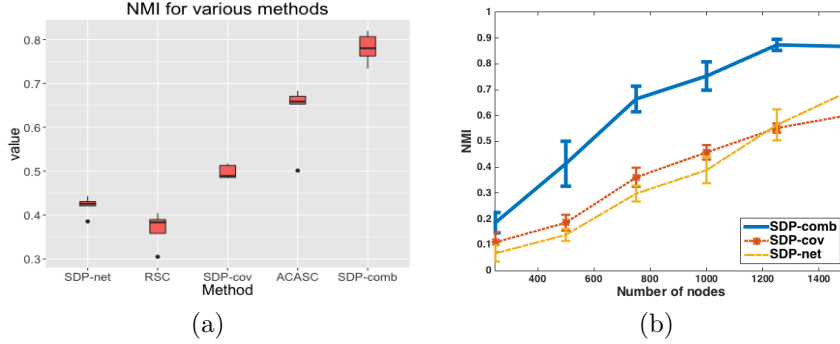


Figure 6: Synthetic Results: (a) Orthogonal sources: $r = 3, 2m_1 = m_2 = m_3, n = 500, \sigma = 1, d = 100$; (b) Asymptotically consistent sources, $r = 3, m_{\max} = 3m_{\min}, B = 0.3 \cdot (0.1E + 0.1I), \min d_{k\ell}^2 = 0.08, \sigma = 1, d = n^{2/3}$.

- [3] Awasthi, P. and Sheffet, O. (2012). Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer.
- [4] Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073.
- [5] Binkiewicz, N., Vogelstein, J. T., and Rohe, K. (2014). Covariate assisted spectral clustering. *arXiv preprint arXiv:1411.2158*.
- [6] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.
- [7] Braverman, M., Makarychev, K., Makarychev, Y., and Naor, A. (2013). The grothendieck constant is strictly smaller than krivine’s bound. In *Forum of Mathematics, Pi*, volume 1, page e4. Cambridge Univ Press.
- [8] Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982.
- [9] Cai, T. T., Li, X., et al. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *The Annals of Statistics*, 43(3):1027–1059.
- [10] Chen, Y., Sanghavi, S., and Xu, H. (2012). Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212.
- [11] Chen, Y. and Xu, J. (2014). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*.
- [12] Dasgupta, S. and Schulman, L. (2007). A probabilistic analysis of em for mixtures of separated, spherical gaussians. *The Journal of Machine Learning Research*, 8:203–226.

- [13] Dhillon, I. S., Guan, Y., and Kulis, B. (2004). Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556. ACM.
- [14] El Karoui, N. et al. (2010). On information plus noise kernel random matrices. *The Annals of Statistics*, 38(5):3191–3216.
- [15] Gil-Mendieta, J. and Schmidt, S. (1996). The political network in mexico. *Social Networks*, 18(4):355–381.
- [16] Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010). A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233.
- [17] Guédon, O. and Vershynin, R. (2015). Community detection in sparse networks via grothendieck’s inequality. *Probability Theory and Related Fields*, pages 1–25.
- [18] Gunnemann, S., Farber, I., Raubach, S., and Seidl, T. (2013). Spectral subspace clustering for graphs with feature vectors. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 231–240. IEEE.
- [19] Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic block-models: First steps. *Social networks*, 5(2):109–137.
- [20] Hsu, D., Kakade, S. M., and Zhang, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.*, 17(52):1–6.
- [21] Jacob, U., Thierry, A., Brose, U., Arntz, W. E., Berg, S., Brey, T., Fetzner, I., Jonsson, T., Mintenbeck, K., Mollmann, C., et al. (2011). The role of body size in complex food webs: A cold case. *Advances In Ecological Research*, 45:181–223.
- [22] Kadavankandy, A., Cottatellucci, L., and Avrachenkov, K. (2015). Characterization of random matrix eigenvectors for stochastic block model. In *2015 49th Asilomar Conference on Signals, Systems and Computers*, pages 861–865. iee.
- [23] Le, C. M., Levina, E., and Vershynin, R. (2015). Sparse random graphs: regularization and concentration of the laplacian. *arXiv preprint arXiv:1502.03049*.
- [24] Lei, J., Rinaldo, A., et al. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237.
- [25] Mixon, D. G., Villar, S., and Ward, R. (2016). Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*.
- [26] Newman, M. and Clauset, A. (2015). Structure and inference in annotated networks. *arXiv preprint arXiv:1507.04001*.

- [27] Oliveira, R. I. (2009). Concentration of the adjacency matrix and of the laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*.
- [28] Overton, M. L. and Womersley, R. S. (1993). Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming*, 62(1-3):321–357.
- [29] Peng, J. and Wei, Y. (2007). Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1):186–205.
- [30] Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915.
- [31] Sarkar, P., Bickel, P. J., et al. (2015). Role of normalization in spectral clustering for stochastic blockmodels. *The Annals of Statistics*, 43(3):962–990.
- [32] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- [33] Von Luxburg, U., Belkin, M., and Bousquet, O. (2008). Consistency of spectral clustering. *The Annals of Statistics*, pages 555–586.
- [34] Xu, Z., Ke, Y., Wang, Y., Cheng, H., and Cheng, J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 505–516. ACM.
- [35] Yan, B. and Sarkar, P. (2016a). On robustness of kernel clustering. *arXiv preprint arXiv:1606.01869*.
- [36] Yan, B. and Sarkar, P. (2016b). On robustness of kernel clustering. *arXiv preprint arXiv:1606.01869*.
- [37] Yang, J., McAuley, J., and Leskovec, J. (2013). Community detection in networks with node attributes. In *Data mining (ICDM), 2013 IEEE 13th international conference on*, pages 1151–1156. IEEE.
- [38] Zhang, Y., Levina, E., and Zhu, J. (2015). Community detection in networks with node features. *arXiv preprint arXiv:1509.01173*.